PRINCIPLES OF NAVAL WEAPON SYSTEMS



CRAIG PAYNE

Principles of Naval Weapon Systems

Titles in the Series

American Admiralship
The Bluejacket's Manual
Career Compass
The Chief Petty Officer's Guide
Command at Sea
Dictionary of Modern Strategy and Tactics
Dictionary of Naval Abbreviations
Dictionary of Naval Terms
Division Officer's Guide
Dutton's Nautical Navigation
Ethics for the Junior Officer
Farwell's Rules of the Nautical Road
Naval Ceremonies, Customs, and Traditions
NavCivGuide: A Handbook for Civilians in the U.S. Navy
The Naval Institute Almanac of the U.S. Navy
The Naval Institute Guide to Naval Writing
The Naval Officer's Guide
Naval Shiphandler's Guide
The Professional Naval Officer: A Course to Steer By
Reef Points
A Sailor's History of the U.S. Navy
Watch Officer's Guide

Forthcoming

Dutton's Navigation Dictionary
The Mariner's Guide to Meteorology
The Mariner's Guide to Oceanography
The Newly Commissioned Officer's Guide
The Petty Officer's Guide
The Submarine Officer's Guide

The U.S. Naval Institute
Blue & Gold Professional Library

For more than 100 years, U.S. Navy professionals have counted on specialized books published by the Naval Institute Press to prepare them for their responsibilities as they advance in their careers and to serve as ready references and refreshers when needed. From the days of coal-fired battleships to the era of unmanned aerial vehicles and laser weaponry, such perennials as the *Bluejacket's Manual* and the *Watch Officer's Guide* have guided generations of Sailors through the complex challenges of naval service. As these books are updated and new ones are added to the list, they will carry the distinctive mark of the Blue & Gold Professional Library series to remind and reassure their users that they have been prepared by naval professionals and meet the exacting standards that Sailors have long expected from the U.S. Naval Institute.



Naval Institute Press
291 Wood Road
Annapolis, MD 21402

© 2006 by the United States Naval Institute
Annapolis, Maryland
All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage and retrieval system, without permission in writing from the publisher.

ISBN 13: 978-1-59114-658-2
Library of Congress Cataloging-in-Publication Data
Principles of naval weapon systems / edited by Lt. Cdr. Craig M. Payne, USN.
p. cm. — (The U.S. Naval Institute Blue & Gold Professional Library)
Includes bibliographical references and index.
ISBN 1-59114-658-5 (alk. paper)
1. Weapons systems. I. Payne, Craig M.
VF346.P755 2006
623.825'1--dc22

2006030946

Printed in the United States of America on acid-free paper: ♾

12 11 10 09 08 9 8 7 6 5 4 3 2

Principles of Naval Weapon Systems

Edited by Lt. Cdr. Craig M. Payne, USN

Revision and new material by
Lt. Ben F. Anton, USN
Lt. Cdr. Nick F. Budd, USN
Maj. Todd W. Ferry, USMC
Capt. Gary D. Noble, USN
Lt. Cdr. William S. Pendergrass, USN
Lt. Cdr. Brian M. Riley, USN
Lt. Gene G. Severston, USN
Lt. Mike J. Supko, USN
Cdr. Timothy P. Tumelty, USN
Capt. Allison D. Webster-Giddings, USN
Lt Cdr. Dan J. Welsh, USN

NAVAL INSTITUTE PRESS
Annapolis, Maryland

Contents

CHAPTER 1 Introduction
Detect to Engage Sequence 1

CHAPTER 2 Electromagnetic Fundamentals
Introduction 5
Characteristics of Traveling Waves 5
Wave Propagation 9
Antennas 15
Electromagnetic Wave Modes of Propagation in the Atmosphere 22

CHAPTER 3 Radar Fundamentals
Introduction 26
Principles of Operation 26
Mechanization 28
Performance Factors 30
The Simplified Radar Range Equation 39
Environmental Considerations 41
Summary 41

CHAPTER 4 Radar Systems
Introduction 42
Continuous-Wave (CW) and Frequency Modulated Continuous Wave (FMCW) Radar 42
Combination Radar Systems 45
High-Resolution Radars 51
Electronic Scanning and the Phased Array 54
Effects of Changing Frequency 63
Summary 64

CHAPTER 5 Tracking Systems
Introduction 66
Servo Systems 66
Track-While-Scan Concepts 75
Summary 82

CHAPTER 6 Electronic Combat
Introduction 83
Electronic Combat 83
Basic Electronic Combat Receivers 83
Electronic Attack 86
Electronic Protection 94
The Effectiveness of EA/EP 98
Summary 98

CHAPTER 7 Principles of Electro-Optics Sensing
Introduction 99
Plane Geometry vs. Solid Geometry 102
Radiometric Quantities 103
Electro-Optical Sources—Targets and Background 108
Thermal Radiation 109
Selective Radiators 114
Reflected (External) Radiation 115
Electro-Optic Attenuation 117
Spreading and Divergence 118

CHAPTER 8 Electro-Optical Targets and Systems
Introduction 125
Targets and Backgrounds 125
Aerodynamic Heating Signature 130
Lasers 132
Optics 134
Scanners and Detectors 137
Simplified IR Range Equation 145
Night Vision Devices (NVDs) 146
Summary 151

CHAPTER 9 Principles of Underwater Sound
Introduction 153
Fundamental Concepts 154
Speed of Sound in the Sea 159
Typical Sound Speed Profiles 162
Propagation Paths 164
Sound Sources and Noise 170

CHAPTER 10 Figure of Merit
Introduction 179
Sonar Equation Terms 179
Passive Sonar Equation 186
Active Sonar Equation 187
Figure of Merit 189

CHAPTER 11 Undersea Warfare Systems
Introduction 192
Transducers 192
Active Systems 196
Passive Systems 200
Variable Depth Sonar 203
Towed Array Sonar System (TASS) 204
Sonobuoys 205
Multistatic Sonar 206
Nonacoustic Detection 207
Fathometer 209
Underwater Communications 209
Countermeasures 210

CHAPTER 12 Introduction to Automatic Control Systems
Introduction 213
Modeling 214
System Time Responses 220
System Response Examples 223
Feedback 227
Block Diagram 228
Open-Loop Transfer Function Example 229
Closed-Loop Transfer Function Examples 230
Error 231
System Controllers or Compensators 233
Composite Example 235
Summary 236

CHAPTER 13 Ballistics and Fire Control
Introduction 237
Relative Motion 237
Interior Ballistics 238
Exterior Ballistics 242
Reference Frames and Coordinate Systems 248
Gyros 250
Coordinate Systems 253
Weapon System Alignment 258
Summary 261

CHAPTER 14 Weapon Propulsion and Architecture
Introduction 262
Launching System 262
Impulse Launchers 262
Reaction Launchers 269
Gravity Launching Systems 272
Functions of a Launching System 272
Weapon Propulsion 276
Impulse Propulsion 276
Reaction Propulsion 276
Gravity-Type Propulsion 282
Weapon Architecture 282
Architecture of Missiles 282
Architecture of Gun Ammunition 289
Architecture of Torpedoes 292
Architecture of Naval Mines 294
Summary 296

CHAPTER 15 Guidance Systems
Introduction 298
Fundamentals of Guidance Systems 298
Phases of Guidance 300
Homing Logic 301
Types of Guidance Systems 304
Summary 312

CHAPTER 16 Fuzing
Introduction 314
Function of the Fuze System 314
Fuze System Classification 317
Proximity Fuzes 317
Nonproximity-Type Fuzes 324
Reliability 326
Probability Theory 327

CHAPTER 17 Chemical Explosives
Introduction 330
Categories of Explosives 330
Characteristics of Military Explosives 332
Measurements of Chemical Explosive Reactions 335

CHAPTER 18 Weapon Damage Effects
Introduction 341
Quantifying Damage 341
Circular Error Probable 342
Warhead Characteristics 345
Blast Warheads 346
Fragmentation Warheads 350
Shaped Charge Warheads 353
Special-Purpose Warheads 355
Underwater Warheads 358

CHAPTER 19 Weapons of Mass Destruction
Introduction 361
Nuclear Weapons 361
Characteristics of Nuclear Explosions 366
Measuring Radiation 372
Nuclear Weapon Damage Prediction 376
Predicting Nuclear Blast Effects 377
Chemical and Biological Weapons (CBW) 382
Chemical Warfare Agents 385
Biological Warfare Agents 389

Sources 393

Index 397

Preface

This text is intended to serve as an introduction to the underlying science and engineering of weapons used in the naval service. The philosophy used in the material selected for this text is that individual weapons come and go, but the principles of their operation largely remain the same. Some subjects are covered in greater detail than needed for an introductory course to allow this text to serve as a basic reference for users to take into professional life.

The text was written to be inclusive of all college majors; as such, a conscious effort was made when possible to apply algebra, geometry, and trigonometry but to avoid calculus. Therefore, many of the equations derived are first order and provide approximations that are sufficient to illustrate the relative performance parameters of variables used in weapon system design. These same theories and principles can then be applied to actual sensors and weapons using operational parameters and specifications determined from technical manuals and warfare publications.

Material has been drawn from previous texts of the same title that have explained the principles for the last forty years. Much of the material can be traced to the work completed by the Bureau of Naval Weapons in the 1960s. It was updated and expanded in the 1980s' version and incorporated into this text. In some cases, principles of systems that the U.S. Navy no longer uses are described in the belief that sometimes it is good to know where you have been in order to know where you are going. In addition, many countries and organizations still employ some of these lower-technology systems. Therefore, it is necessary to understand their basic capabilities. With the advent of new technologies and methods, this text will require periodic updating.

Work of this magnitude cannot be undertaken by one person alone. Many people over the years are responsible for the scholarship that is incorporated in this text. The U.S. Naval Academy Weapons staff of the early twenty-first century built on the effort put forth by LCDR David Frieden and staff of the 1980s by updating the material.

Specific contributors to this edition of *Principles of Naval Weapon Systems* include LCDR Brian Riley—explosives and weapons of mass destruction; MAJ Todd Ferry and CAPT Allison Webster-Giddings—electro-optics; CAPT Owen Thorp and LCDR Curt Lindsay—control systems; CAPT Jack Nicholson—underwater sound; CDR Tim Tumelty—electronic combat. Other members of the staff who served as readers offering suggestions and providing corrections include LT Gene Severtson, LT Ben Aton, LCDR Mike Supko, CAPT Gary Noble, LCDR Jennifer Nichols, LCDR Stan Pendergrass, and LCDR Nick Budd. Also, special mention should be made of LCDR Dan Welsh, NROTC University of Notre Dame, who provided input to ensure that all midshipmen, regardless of commissioning source, are provided the same material.

Previous students from the USNA classes of 2003, 2004, and 2005 need to be acknowledged as well for reading two draft versions as part of class and finding errors or omissions in the content, but most important my spelling and grammatical errors.

The Naval Institute's senior acquisitions editor, Tom Cutler, was invaluable in helping to keep the production on this text moving forward. Much appreciation must go to this very professional publisher.

1 Introduction

The intent of this text is to provide a broad overview of the basic principles of weapon systems used in the naval service. It is not desired or intended to be a study of specific weapon systems but instead to be a presentation of how a weapon system interacts within the physical constraints of the environment. This approach relies on the view that, although technology progresses and new weapon systems are built, these systems must ultimately interconnect the weapon with the target in the environment in order to be effective. With this in mind, specific examples of weapon systems are given to show real-world applications of this coupling.

The principles addressed in this text are drawn from diverse practical experience and a wide range of academic disciplines to provide present and future officers with a basic foundation in weapon systems. By no means should this text be considered all-encompassing or the definitive word concerning the principles addressed. Rather, the reader should view it as a basis for subsequent functional training and continuing education in the art and science of naval weapon systems and their applications.

Detect to Engage Sequence

The international situation has deteriorated and the United States and Nation X have suspended diplomatic relations. The ruler of Nation X has threatened to annex the smaller countries bordering Nation X and has threatened hostilities toward any country that tries to stop him. You are assigned to a guided missile cruiser that is a member of Battle Group Zulu, currently stationed approximately 300 nautical miles off the coast of Nation X. The Battle Group Commander has placed the Battle Group on alert by specifying the Warning Status as Yellow in all warfare areas, meaning that hostilities are probable.

You are standing watch as the Tactical Action Officer (TAO) in the Combat Information Center (CIC), the nerve center for the ship's weapon systems. Dozens of displays indicate the activity of ships and aircraft in the vicinity of the Battle Group. As the TAO you are responsible for the proper employment of the ship's weapon systems in the absence of the Commanding Officer. It is 0200 and you are in charge of a multimillion-dollar weapon system and responsible for the lives and welfare of your shipmates.

The relative quiet of the CIC is shattered by an alarm on your Electronic Combat (EC) equipment indicating the initial ***detection and identification*** by your Electronic Support (ES) equipment of a possible incoming threat. The wideband ES receiver detects an electromagnetic emission on a bearing in the direction of Nation X. Almost instantaneously the emitter's parameters are interpreted by the equipment and compared with radar parameters stored in the memory of the ES equipment. On a display screen a symbol indicating the emitter's approximate line of bearing from your ship is presented. You notify the Commanding Officer of this new development. Meanwhile, the information is transmitted to the rest of the Battle Group via radio data links.

Moments later, in another section of the CIC, the ship's long-range two-dimensional air-search radar is just beginning to pick up a faint return at the radar's maximum range. The information from the air-search radar coupled with the line of bearing from your ES equipment allows you to ***localize*** the contact and determine an accurate range and bearing. Information continues to arrive, as the ES equipment ***classifies*** the J-band emission as belonging to a Nation X attack aircraft capable of carrying antiship cruise missiles.

The contact continues inbound, headed toward the Battle Group, and within minutes is within range of your ship's three-dimensional search-and-track radar. The contact's bearing, range, and altitude are plotted to give an accurate course and speed. The range resolution of the pulse-compressed radar allows you to determine that

the target is probably just one aircraft. You continue to ***track*** the contact as you ponder your next move.

As the aircraft approaches the outer edge of the range of its air launched cruise missile (ALCM), the ES operator reports that the aircraft's radar sweep has changed from a search pattern to a single-target track mode, indicating imminent launch of a missile. In accordance with the Rules of Engagement (ROE) in effect, you have determined that this action shows hostile intent on the part of the target, and decide to defend the ship against imminent attack. You inform your CIC team of your intentions, and ***select a weapon***, in this case a surface-to-air missile, to engage the target. You also inform the Air Defense Commander of the indications of hostile intent, and he places you and the other ships in Air Warning Red, attack in progress.

As the target closes to the maximum range of your weapon system, the fire control or tactical computer program, using target course and speed, computes a predicted intercept point (PIP) inside the missile engagement envelope. Your control screen indicates that the PIP is within the engagement envelope and that your weapon system is "locked on" target. You authorize batteries released and the missile is launched toward the PIP. As the missile speeds toward its target at Mach 2+, the ship's sensors continue to track both the aircraft and the missile. Guidance commands are sent to the missile to keep it on course.

On board the enemy aircraft, the pilot is preparing to launch an ALCM, when his own ES equipment indicates that he is being engaged. This warning comes with but precious few seconds, as the missile enters the terminal phase of its guidance. In a desperate attempt to break the radar lock, he employs evasive maneuvering. It's too late, though, as the missile approaches the lethal "kill radius," and the proximity fuze on the missile's warhead detonates its explosive charge, sending fragments out in every direction, ***destroying or neutralizing*** the target. This information is confirmed by your ship's sensors. The radar continues to track the target as it falls into the sea and your ES equipment goes silent.

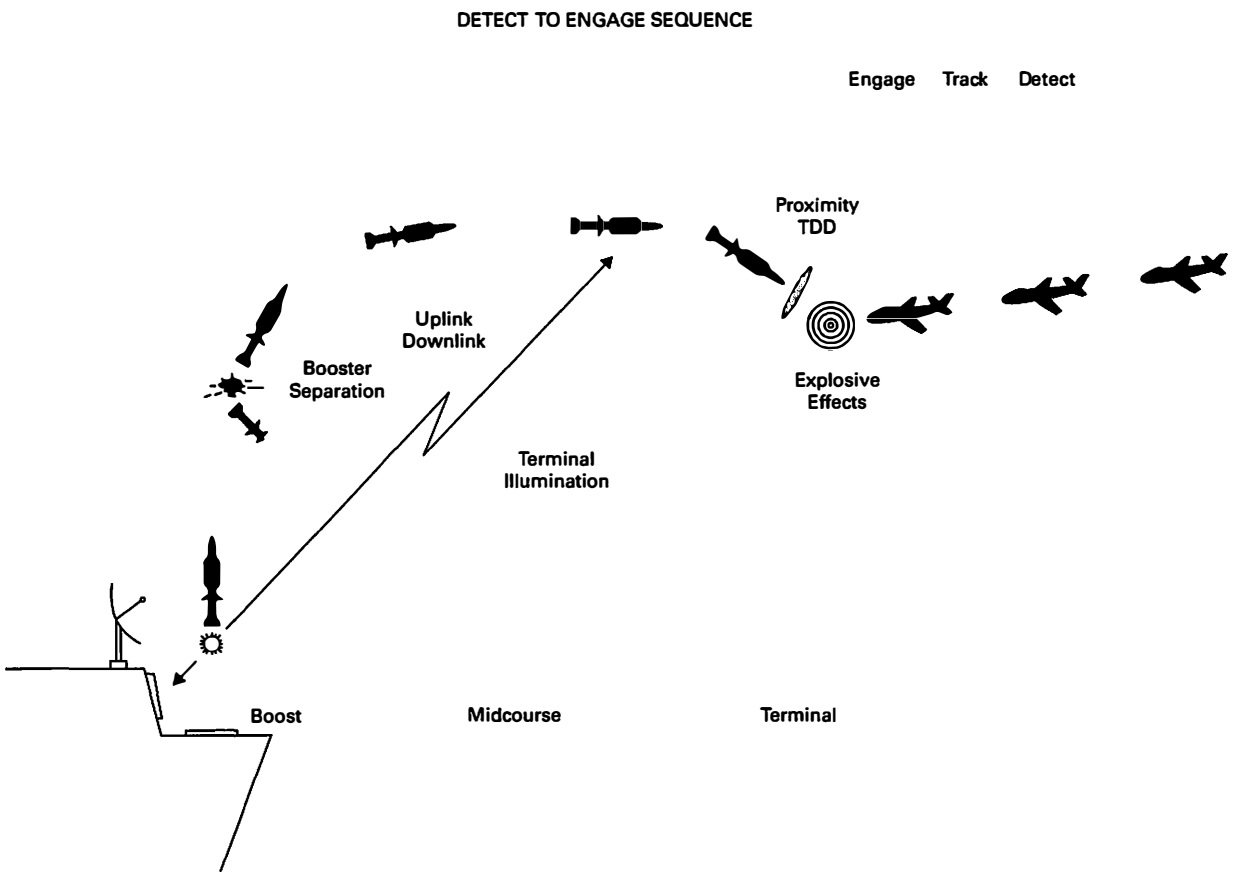


FIGURE 1-1 *Typical shipboard detect to engage sequence*

The Fire Control Problem

What has just been described is not something out of a war novel, but rather a scenario of a possible engagement between a hostile force (the enemy attack aircraft) and a Naval Weapon System (the ship). This scenario illustrates the concept of the "detect to engage" sequence, which is an integral part of the modern fire control problem; see figure 1-1. Although the scenario was one of a surface ship against an air target, every weapon system performs the same functions: ***target detection, resolution or localization, classification, tracking, weapon selection, and ultimately neutralization***. In warfare, these functions are accomplished by submarines, aircraft, tanks, and even marine infantry. The target may be either stationary or mobile; it may travel in space, through the air, on the ground or surface of the sea, or even beneath the sea. It may be manned or unmanned, guided or unguided, maneuverable or in a fixed trajectory. It may travel at speeds that range from a few knots to several times the speed of sound.

The term *weapon system* is a generalization encompassing a broad spectrum of components and subsystems. These components range from simple devices operated manually by a single person and accomplishing one specific function, to a complex array of subsystems interconnected by computers and data communication links capable of performing several functions or engaging numerous targets simultaneously. Although each subsystem may be specifically designed to solve a particular part of the fire control problem, it is these components operating in concert that allows the whole system to achieve its ultimate goal—the neutralization of the target.

Components

All modern naval weapon systems, regardless of the medium they operate in or the type of weapon they employ, consist of the basic components that allow the system to detect, track, and engage the target. Sensor components must be designed for the environments in which the weapon system and the target operate. These components must also be capable of coping with widely varying target characteristics, including target range, bearing, speed, heading, size, and aspect.

Detecting the Target

There are three phases involved in target detection by a weapon system. The first phase is surveillance and detection, the purpose of which is to search a predetermined area for a target and detect its presence. This may be accomplished actively, by sending energy out into the medium and waiting for the reflected energy to return, as in radar; or passively, by receiving energy being emitted by the target, as in our scenario.

The second phase is measuring or localizing the target's position more accurately and, by a series of such measurements, estimating its behavior or motion relative to the detecting platform. This is accomplished by repeatedly determining the target's range, bearing, and depth or elevation.

Finally, the target must be classified, that is, its behavior must be interpreted so as to estimate its type, number, size, and, most important, identity. The capabilities of weapon system sensors are measured by the maximum range at which they can reliably detect a target and its ability to distinguish individual targets in a multitarget group. In addition, sensor subsystems must be able to detect targets in a medium cluttered with noise, which is any energy sensed other than that attributed to a target. Such noise or clutter is always present in the environment due to spurious reflections or as a result of deliberate interference or jamming. Noise is also self-generated, for example within the electronic circuitry of the detecting device itself.

Tracking the Target

Sensing the presence of a target is an essential first step to the solution of the fire control problem. To successfully engage the target and solve the problem, updates as to the target's position and its velocity relative to the weapon system must also be known or estimated continuously. This information is used not only to evaluate the threat represented by the target, but also to predict the target's future position along with a weapon intercept point so that the weapon can be accurately aimed and controlled. The fire control solution is the convergence of the target's predicted position with the weapon intercept point.

In order to obtain target trajectory information, methods must be devised to enable the sensor to follow or track the target. This control or "aiming" may be accomplished by a collection of motors and position-sensing devices called a servo system with associated feedback loops. Another method involves stationary sensing devices that are electronically steered in the direction of the target using interference properties of waves.

The scenario presented in the beginning of this chapter was in response to a single target. In reality, this is rarely the case. The modern "battlefield" is one in which sensors are detecting numerous contacts—friendly, neutral, and hostile—and information is continually being gathered on all of them. The extremely high speed, precision, and flexibility of modern computers enable

the weapon systems and their operators to compile, coordinate, and evaluate the data, and then initiate an appropriate response. Computers enable a weapon system to detect, track, and predict target motion automatically. This establishes the target's presence and defines how, when, and with what weapon the target will be engaged.

Engaging the Target

Effective engagement and neutralization of the target requires that a destructive mechanism, in this case a warhead, be delivered to the target's vicinity or actually impact the target. How close to the target a warhead must be delivered depends on the type of warhead and the type of target. In delivering the warhead, the aiming, launch, type of propulsion system of a weapon, and the forces to which the weapon are subjected en route to the target need to be considered. The weapon's capability to be guided or controlled after launch dramatically increases its accuracy and probability of kill from a single weapon. The use of guidance systems also dramatically complicates system designs. These factors as well as the explosive to be used, the fuzing mechanism, and warhead design are all features in the design and effectiveness of a modern weapon.

Conclusion

Solving the fire control problem, from target detection to neutralization, requires a complex integration of numerous components that must operate together to achieve the goal. At the functional center of every weapon system is a human being who is engaged in the employment of that system. These systems require that a U.S. Navy/ Marine Corps officer be competent in the scientific and engineering principles that underlie their design, operation, and employment. Understanding the basic rules for evaluating or predicting a weapon system's performance is necessary to properly employ the system in combat, recognize its limitations and capabilities, and use it as an aid to the development of imaginative tactics. Unless every weapon system operator possesses a thorough understanding of the physical processes that define the performance of equipment, our operating forces will be unable to make full use of their assets and exploit enemy weaknesses and limitations.[i]

---

[i] *FAS*, intro.

2 Electromagnetic Fundamentals

Introduction

A primary requirement of a weapon system is that it have some means of detecting a target. To accomplish detection, the weapon system must be capable of sensing some unique characteristic that differentiates or identifies a target from the background. One such characteristic is the energy that is either emitted or reflected by the target itself. This energy may be in several forms, including electrical, audio, heat, or visible light. A property common to all the energy forms listed above is their manner of propagation. That is, they all propagate in the form of traveling waves and therefore can be defined and categorized by their frequency and wavelength. It is the function of the sensor system, then, to detect the appropriate energy form and to furnish the information obtained to the other components of the weapon system.

This chapter is heavily directed toward the properties of radar energy, but the principles presented here are the same for all other types of electromagnetic energy—such as radio, infrared, visible light, and X rays. Sound energy, though not electromagnetic (EM), also exhibits many of these same properties, as will be described later in this text. It is imperative that a solid foundation in the principles of energy transmission and propagation be acquired in order to fully understand the design and use of various types of modern weapon-system sensors.

Characteristics of Traveling Waves

Energy moves from a source in waves, in much the same way that waves spread out concentrically from the point of impact of a pebble dropped in water. Although waves created by this familiar example are two-dimensional, electromagnetic energy radiated from a point in a vacuum travels as three-dimensional waves—that is, concentric spheres. In the study of radiated energy, it's often difficult to describe expanding spheres of concentric wave fronts as they propagate through space. It is convenient, therefore, to trace the path of rays rather than the waves themselves. A ray is formed by tracing the path of a hypothetical point on the surface of a wave front as it moves through a medium.

Frequency

The principal distinguishing characteristic of any form of radiated energy is its frequency, which is normally measured in cycles per second or hertz (Hz). In the case of sound, it is the rate at which an object vibrates. For electromagnetic energy such as radio, radar, and light, it is the rate at which the electric and magnetic fields of a propagating wave expand and collapse. Since radiated energy is composed of periodic waves, it can be represented by a sinusoid or sine wave. The significance of frequency is clearly evidenced by the fact that the nature of the energy form (i.e., heat, light, electromagnetic, etc.) and its frequency are directly interrelated.

Figure 2-1 shows the frequency spectrum from audio frequencies to gamma rays. Notice that radar energy is

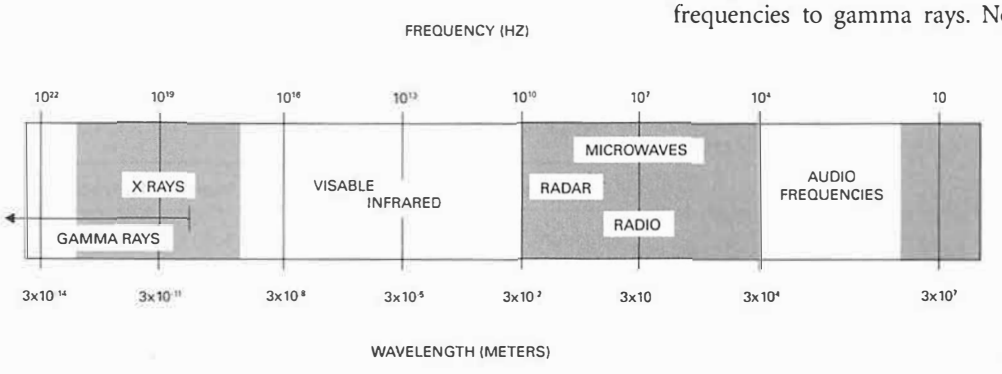


FIGURE 2-1 *Frequency spectrum*

normally between 1 GHz and 10 GHz, but there are some exceptions. For example, some long-range air-search radars transmit between 200 MHz and 980 MHz, and over-the-horizon radars transmit at even lower frequencies.

Wavelength

Wave motion is further described by wavelength (λ); see figure 2-2. Wavelength is defined as the distance between two identical points on adjacent waves (normally, crests). For a traveling wave, it is also a measure of the distance covered by the wave during one complete cycle. From the latter definition, a relationship between frequency and wavelength is:

$$\lambda = \frac{v}{f} \qquad (2\text{-}1)$$

where:

λ = wavelength (meters)
v = wave propagation velocity in a particular medium (meters/sec)
f = frequency (Hz)

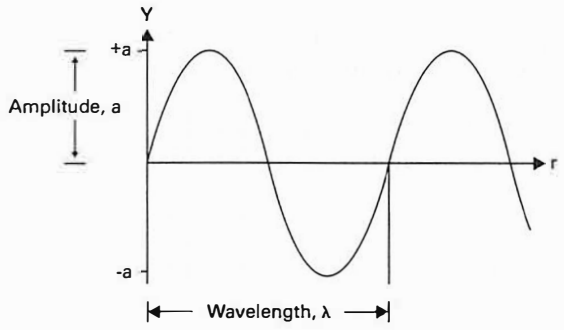


FIGURE 2-2 *Parts of a sine wave*

The relationship of equation 2-1 describes how the speed of propagation varies with frequency. For instance, a prism disperses light into separate colors (frequencies) because they propagate at different speeds. Components with lower speeds of propagation bend by a larger amount when they enter a new medium in accordance with Snell's Law, (discussed later). This causes the different frequency components (colors) to be separated or dispersed; see figure 2-3.[i]

As will be shown in the next chapter, wavelength is important in the design of radio and radar antennas, since for any given amount of signal, directivity, and size of the antenna is directly proportional to the wavelength of the energy being directed.

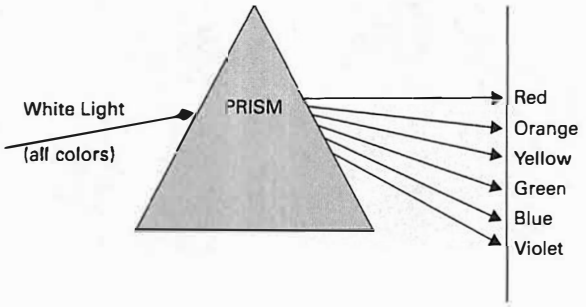


FIGURE 2-3 *Prism dispersion of white light*

Coherency

Any pure energy of singular frequency will have the waveform of a sinusoid and is termed *coherent* energy. Energy that is composed of more than one frequency will have some waveform other than a sinusoid and is considered *noncoherent* energy.

Velocity

Electromagnetic energy propagates through a vacuum at the speed of light, approximately 3×10^8 m/s. This equates to approximately 1 nautical mile every 6.18 μs. The velocity of light in other media is less than that in a vacuum.

The ratio of the velocity of electromagnetic energy in a vacuum to the velocity in a different medium is called the index of refraction of that medium. In media such as water or glass, the speed of propagation is slower. If the speed of light in a vacuum is given the symbol c, and its speed in some medium is v, the index of refraction, n, is then defined as:

$$n = c/v \qquad (2\text{-}2)$$

For common media, as shown in table 2-1, the index of refraction is always greater than 1, meaning the speed of propagation is less than that in a vacuum. At radar frequencies and above, the index of refraction in air is very close to 1.0 and therefore is usually ignored.

TABLE 2-1 *Indexes of refraction for common substances*

Substance	Index of Refraction
Vacuum	1
Air	1.0003
Water	1.33
Glass	1.55

Amplitude

Another characteristic of wave propagation is amplitude. This may be graphically defined as the maximum displacement of any point on the wave from a constant reference value. Displacement is also a direct indication of the level of energy at any point on a propagating wave. In order to derive an equation that completely defines displacement in terms of time and distance, it is helpful to begin by investigating the simple example of a standing sine wave at a single instant in time; see figure 2-4.

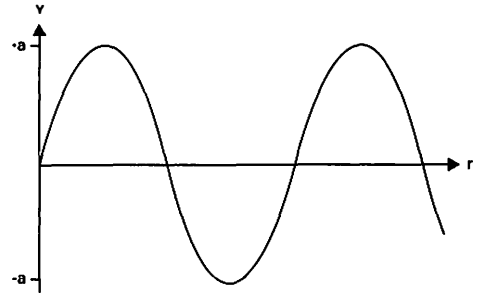


FIGURE 2-4 *Standing sine wave*

The y displacement of this waveform at any given point along the r axis is:

$$y = a\sin\frac{2\pi r}{\lambda} \qquad (2\text{-}3)$$

The y displacement is solely a function of the lateral (r) displacement from the origin since the above waveform is fixed with respect to time.

Now consider a wave that is propagating to the right. To envision this effect, first imagine that the waveform of figure 2-4 is moving to the right at a constant velocity, and then focus on the resulting variation in displacement at some fixed distance along the r axis. Figure 2-5 illustrates

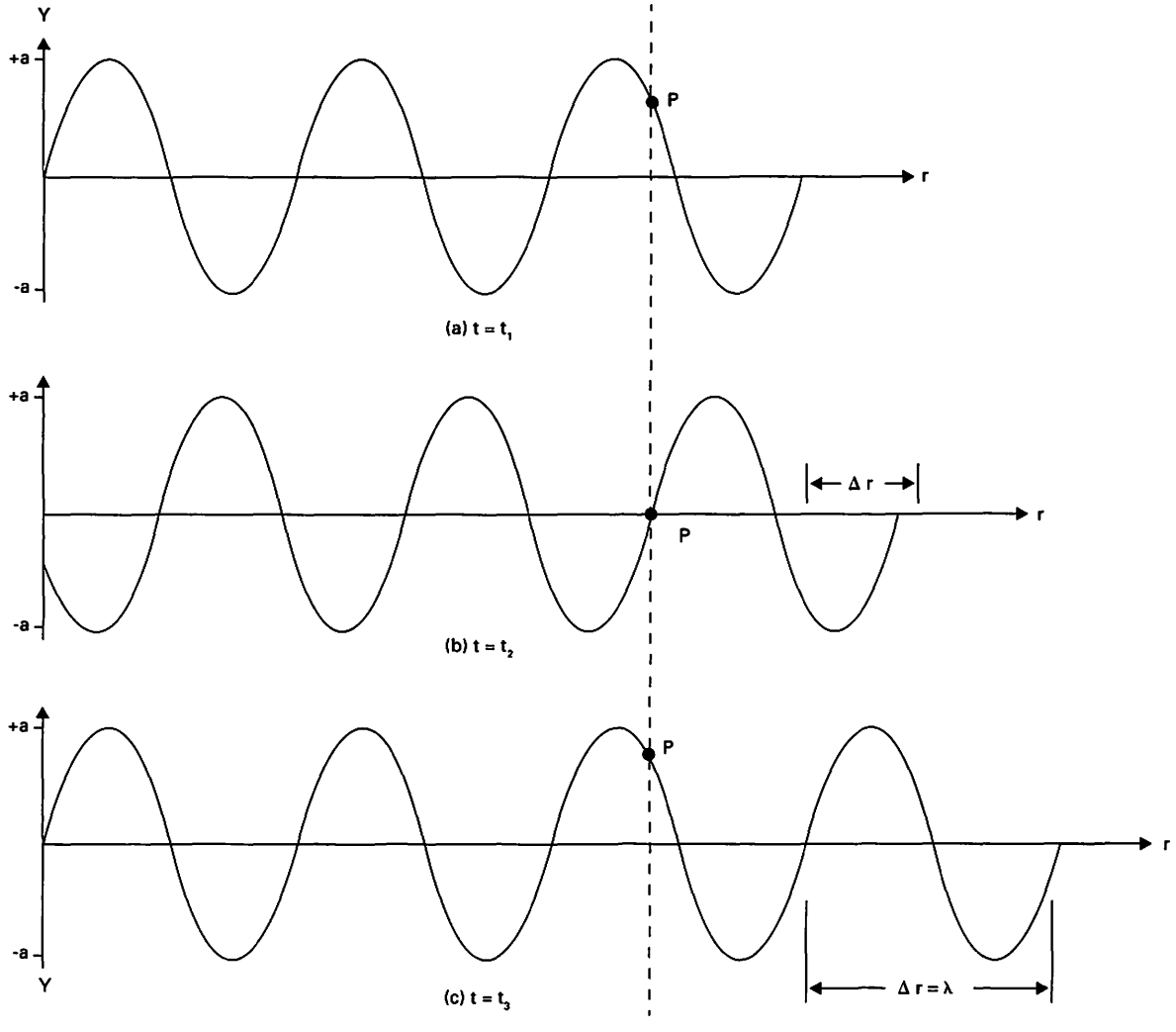


FIGURE 2-5 *Propagating wave*

this variation. In it, the displacement of a point (P) at a fixed lateral position is observed at three successive time intervals.

At time $t = t_1$, displacement of the wave at point P is shown in figure 2-5a. This displacement will vary as the wave propagates in direct proportion to the frequency of oscillation. Figure 2-5b represents some later time, t_2. Finally, time t_3 is chosen such that during the interval $t_3 - t_1$ the distance (Δr) traveled by the wave is equal to one wavelength ($\Delta r = \lambda$). This is illustrated in figure 2-5c. Note that point P has returned to its original position and has thus undergone a complete cycle. The time required for a full cycle is termed the period T, and is mathematically determined as the inverse of frequency:

$$T = 1/f \tag{2-4}$$

where:

T = period (sec)

The number of cycles through which any point will have passed during an elapsed time (t) is equal to the frequency multiplied by elapsed time ($f \cdot t$). Since a full cycle is equivalent to 360° or 2π radians, the phase angle of any point after an elapsed time (t) is $2\pi ft$.

It is now possible to express the displacement at any point on a propagating wave as a function of distance from the source, elapsed time, and amplitude:

$$y = a\sin\left(\frac{2\pi r}{\lambda} - 2\pi ft\right) \tag{2-5}$$

and since

$$\lambda = \frac{v}{f} \quad \text{or} \quad f = \frac{v}{\lambda} \tag{2-6}$$

the equation becomes:

$$y = a\sin\left(\frac{2\pi r}{\lambda} - 2\pi\frac{v}{\lambda}t\right)$$

or (2-7)

$$y = a\sin\left(\frac{2\pi}{\lambda}(r - vt)\right)$$

where:

y = displacement in the y direction
a = amplitude
λ = wavelength (meters)
r = distance from wave origin (meters)
v = velocity of propagation (meters/sec)
t = elapsed time (sec)

Phase

The last parameter necessary to completely describe a wave, but which cannot be readily understood from figure 2-5, is the phase of the wave. Phase is the angular argument (in degrees or radians) at which the sine wave begins, as defined relative to some arbitrary starting point. The symbol for phase is either ϕ or θ. Phase shift is sometimes represented by $\Delta\phi$ or $\Delta\theta$ (the Greek symbol Δ, capital "delta," is used to denote change in a quantity).

Examine figure 2-6 and notice that these two waves have the same wavelength and amplitude. Both have the same basic shape; however, they differ in the starting point. This difference is called the phase shift. For a single wave, the phase shift is undefined because it corresponds to some arbitrary starting point. In contrast, the starting point between two or more waves is important; therefore the relative phase between two or more waves is needed to fully describe the relationship between them.

The concept of phase shift is significant and requires some care to quantify it. The direction of a positive and negative phase shift must be defined. In figure 2-6, wave 2 appears to be advanced relative to wave 1; by convention this is considered a positive phase shift. Another way to visualize this is to move the entire figure of wave 2 to the left for a positive phase shift. To avoid confusion, make the following association:

positive phase shift ⇔ advanced wave
negative phase shift ⇔ retarded wave

As with any physical quantity, the phase shift must have units, and the units must represent this shift in one wave, as compared to the reference wave. Phase shift therefore has units of an angle, because the sine wave is based on a circle and repeats every 360°. Therefore phase shifts range between 0° and 360°. A phase shift of 0° represents no shift, and 360° corresponds to a shift of exactly one wavelength, which is again equivalent to no shift at all.

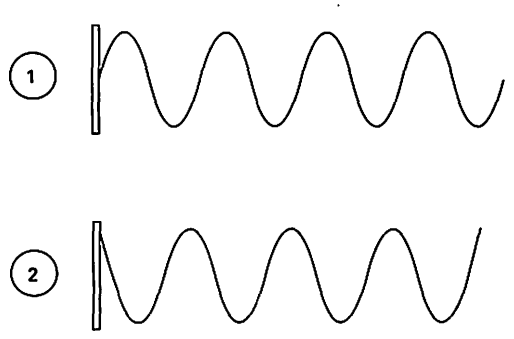


FIGURE 2-6 *Phase difference between two waves*

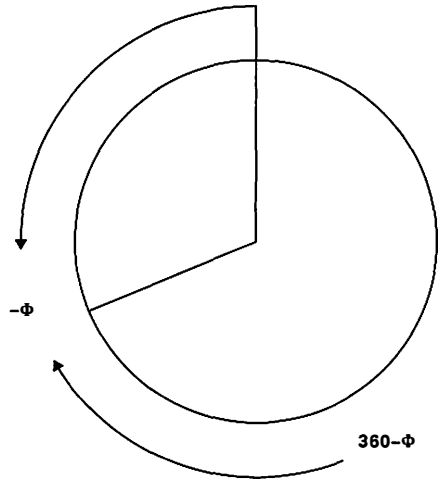


FIGURE 2-7 *The equivalence of positive and negative phase shifts*

Also note that a negative phase shift of $-\phi$ degrees is the same as the positive phase shift of $360° - \phi$. These are just two equivalent ways to describe the direction on a circle, one counterclockwise and the other clockwise.[ii]

Units of phase are sometimes expressed as the unitless radian. Since there are 2π radians in a 360° circle, use the following relationship to convert between degrees and radians:

$$\text{radians} = (2\pi/360°) \times \text{degrees}$$

$$\text{degrees} = (360°/2\pi) \times \text{radians}$$

This conversion factor is approximately 57.3° per radian. The concept of phase shift is essential to the operation of antennas and sonar arrays.

Frequency vs. Time Domain

When plotting the wave's amplitude as a function of time (at a fixed location), the time domain description is used, meaning that the horizontal axis has units of time. The wave has parameters of amplitude, period, and phase. An alternate and completely valid way to describe the same wave is to specify the amplitude and frequency.

The description of a wave at points in time is the representation in the time domain. Description of the wave at each frequency (for sine and cosine waves) is the representation in the frequency domain. They should be thought of as two completely equivalent methods of representing the same thing. Examine the simple sine wave in both representations in figure 2-8. In the time domain, the sine wave completes eight cycles in 1 second. In the frequency domain, there is a single component at 8 Hz.[iii]

Wave Propagation

Basic Phenomena

Energy moves from a source in waves, in much the same way that waves spread out concentrically from the point of impact of a pebble dropped in water. While waves created by this familiar example are two-dimensional, electromagnetic waves radiating from a point in a vacuum travel as three-dimensional waves—that is, concentric spheres; figure 2-9.

In the study of radiated energy, it's often difficult to describe expanding spheres of concentric wave fronts as they propagate through space. It is convenient, therefore, to assume that as the wave spreads out far from its origin, the wave will appear to have the same amplitude everywhere

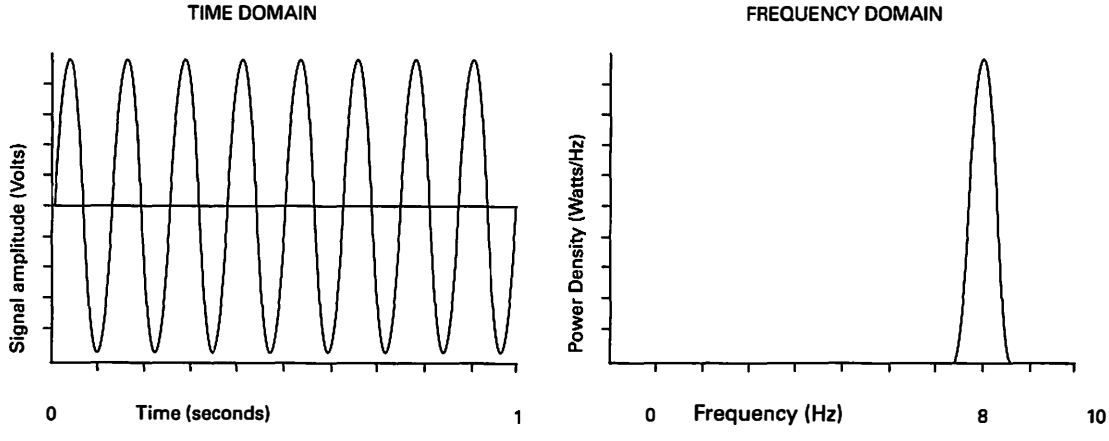


FIGURE 2-8 *Time and Frequency Domain representation of cosine wave*



FIGURE 2-9 *Cross section of spherical wave spreading*

on the plane perpendicular to the direction of travel. This type of wave is called a plane wave. A plane wave is an idealization that allows an approximation of the entire wave traveling in a single direction, instead of spreading out spherically; see figure 2-10.

Reflection

When an electromagnetic wave encounters a conducting surface, reflection of energy from that surface occurs. This reflection is in accordance with the law stating that the reflected and incident waves travel in directions that make equal angles with the perpendicular to the reflecting surface and are in the same plane; see figure 2-11. The angles are called angle of reflection and angle of incidence, respectively, and are measured from the normal (perpendicular) to the reflecting surface. Reflection from a smooth surface as perceived by the wave is called specular

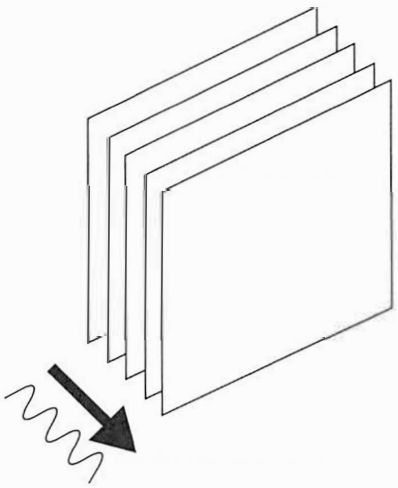


FIGURE 2-10 *Plane wave spreading*

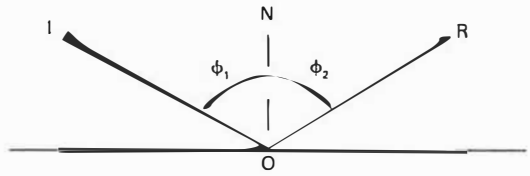


FIGURE 2-11 *Reflection where the incident ray is IO and the reflected ray is OR. The angle of incidence, ϕ_1, equals the angle of reflection, ϕ_2*

reflection. An example of this type of reflection is light off a mirror. ("Speculum" is Latin for mirror.) Uneven surfaces reflect light in a multitude of directions and this is called diffuse reflection.

Reflection can be expressed in terms of the reflection coefficient of a surface, that is, the ratio of the intensities of the reflected field to the incident field. The most common measure of intensity is defined by the electric field strength. Often energy is lost from the incident wave because of the presence of natural obstacles such as dust, snow, or water vapor. These will cause some degree of diffuse reflection, resulting in a loss in energy in the desired direction. The greatest loss in field strength during reflection, however, occurs from the diffusion caused by the roughness and irregularities in the conducting surfaces themselves; see figure 2-12.

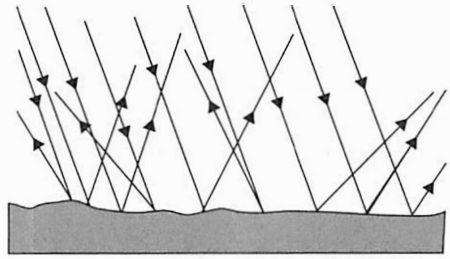


FIGURE 2-12 *Diffusion from an irregular surface*

There is potentially a phase shift to the wave upon reflection as well. Figure 2-13 shows two waves reflected from the earth's surface, like light off a mirror. Notice that the positive and negative alternations of radio waves A and B are in phase before they strike the earth's surface. However, after striking a solid reflecting surface, the radio waves are approximately 180° out of phase.

The amount of phase shift that occurs is not constant. It varies, depending on the wave orientation, the angle at which the wave strikes the surface, and the relative motion between the source and the reflecting surface. Because this reflection is not constant, fading occurs. Normally, radio waves reflected in phase produce stronger signals, whereas those reflected out of phase produce a weak or fading signal.

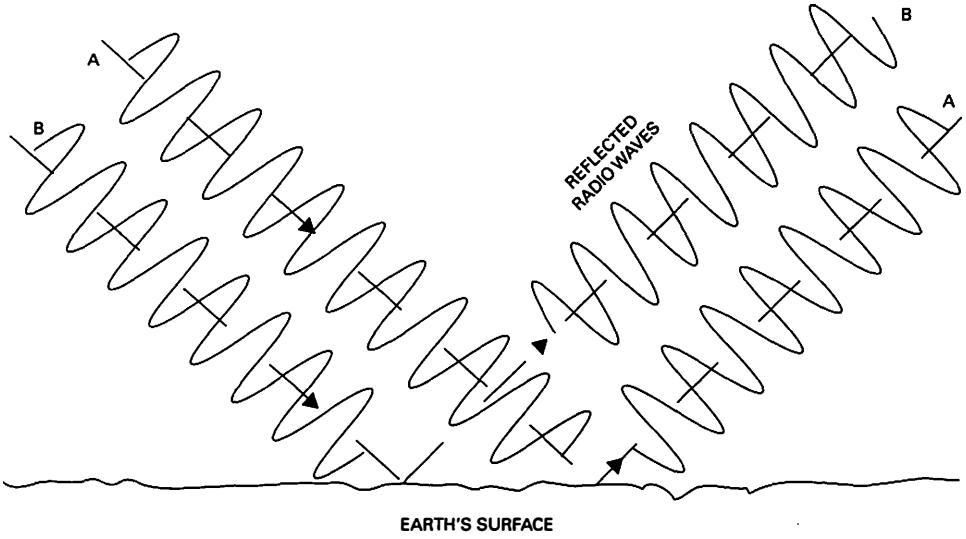


FIGURE 2-13 *Phase shift on reflection*

Refraction

As an electromagnetic wave travels through the atmosphere, it is also subject to the phenomenon of refraction. When a wave strikes a boundary between two transparent media in which the velocity of light differs, the incident ray will generally divide into a reflected ray and a refracted ray. For example, if the new medium has a higher index of refraction, the wavelength must be shorter (the frequency must stay the same due to boundary conditions). The direction of propagation in the new medium must be closer to the normal. The angle of transmission, shown in figure 2-14, will be less than the angle of incidence. Snell's Law predicts the relationship between the angles of incidence and transmission that depend only on the relative indices of refraction:

$$n_1 \sin\phi_1 = n_2 \sin\phi_2 \qquad (2\text{-}8)$$

where:

n_1, n_2 = index of refraction in their respective medium
ϕ_1 = angle of incidence in medium 1
ϕ_2 = angle of refraction in medium 2

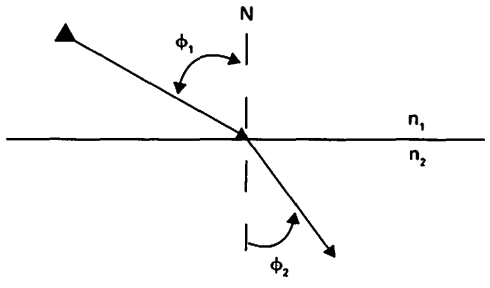


FIGURE 2-14 *Refraction*

While there are no clearly defined media "boundaries" in the earth's atmosphere, an electromagnetic wave will encounter several variations in air density. The index of refraction of air increases with density, and thus there is a downward bending of the rays of propagation. One advantage of such bending of the radar waves is the extension of the radar horizon slightly beyond the absolute straight line-of-sight.

Diffraction

Diffraction causes plane waves traveling in a straight path to bend around a boundary or obstruction. This may be observed at the end of a breakwater or a solid pier when ocean waves approach from an acute angle to the structure. The waves may be observed to bend around the end of the breakwater, resulting in a reduction of the protected area shoreward from it; see figure 2-15.

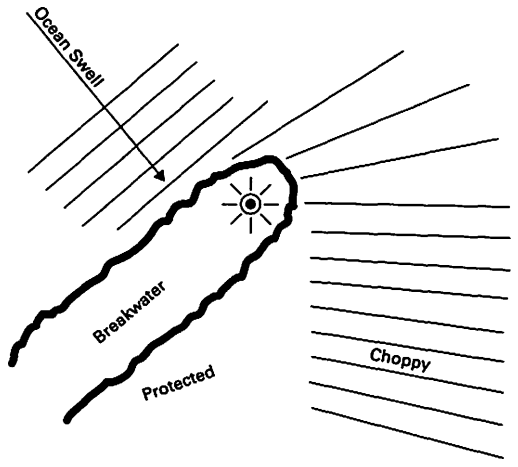


FIGURE 2-15 *Diffraction of ocean swell around a breakwater*

Similar effects can be observed in electromagnetic waves when they encounter an obstacle is such as the edge of an antenna or even large objects such as a mountain or a island, depending on the wavelength. This phenomenon accounts for some losses in antennas and the fact that, depending on the geometry, radar contact or radio communication can sometimes be achieved behind an object; see figure 2-16.

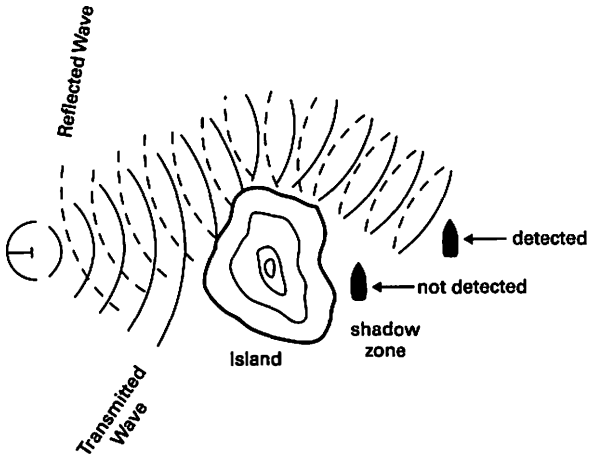


FIGURE 2-16 *Diffraction of a radar wavefront around an obstruction*

Diffraction also refers to the phenomenon of a wave passing through an opening called an aperture, to spread out from that opening, thus creating the basis of antenna beam theory, discussed later in the text. The amount the wave spreads out depends upon the size of the aperture opening relative to the wavelength. When the aperture is large compared to the wavelength, there will be no effect on the wave and it therefore will not diffract. In this case the wave will pass directly through the opening, but the sides will be clipped to match the size of the opening. When the opening is very small compared to the wavelength, the wave will act as if it were again at its origin. In this situation, the wave spreads out uniformly in all directions from the aperture center; see figure 2-17.

The amount of diffraction can be calculated. As an example, consider a circular aperture. If a wave with wavelength λ encounters an opening with diameter D, the amount of diffraction measured by the angle θ (in radians) at which the new wave diffracts from the opening, measured from one edge to the other, or null to null, is $\theta \approx 2.4\lambda/D$. If the opening was a vertical slit, the factor in front would be 2.0 instead of 2.4. The exact value of this factor is some value on the order of 2, and varies with the exact shape of the aperture opening.[iv]

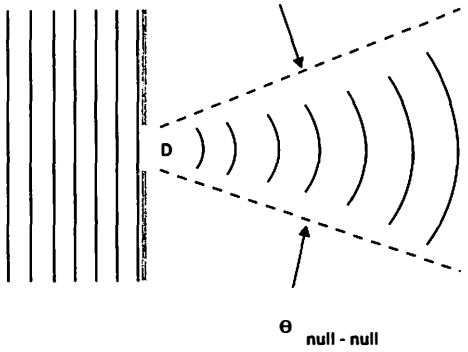


FIGURE 2-17 *Diffraction through a slit*

Interference

All electromagnetic waves can be superimposed upon each other without limit. The electric and magnetic fields simply add at each point. If two waves with the same frequency are combined, there will be a constant interference pattern caused by their superimposition. This interference can either be constructive, meaning that the overall field strength has increased, or destructive, meaning that the field strength has decreased. The type of interference at any point depends on the phase difference between the two waves at that point.

Consider two sources, each of which is operating at the same frequency, but located such that one antenna is a quarter of the wavelength farther away from a common reference point, as shown in figure 2-18. Remember that a path difference of one wavelength corresponds to a phase change of 2π. For simplicity, only the wave's electric field from each antenna is represented.

The field strength at point A from source 1 is given by:

$$E_1 = E_0 \sin\left[\frac{2\pi}{\lambda}(r - ct)\right] \qquad (2\text{-}9)$$

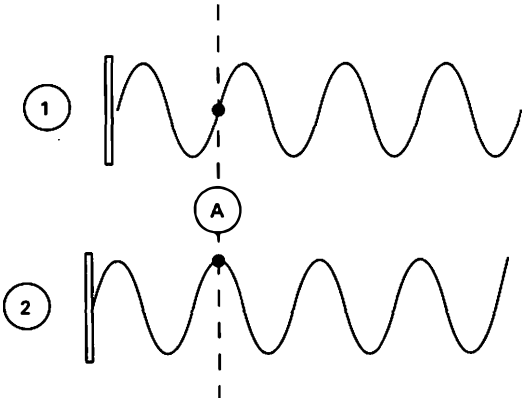


FIGURE 2-18 *Phase difference due to distance from source*

The field strength at point A from source 2 (located $\lambda/4$) from antenna 1 is given by:

$$E_2 = E_0 \sin\left[\frac{2\pi}{\lambda}\left(\left(r+\frac{\lambda}{4}\right)-ct\right)\right] \qquad (2\text{-}10)$$

which can be reduced to

$$E_2 = E_0 \sin\left[\frac{2\pi}{\lambda}(r-ct)+\frac{\pi}{2}\right] \qquad (2\text{-}11)$$

where $\pi/2$ is the phase angle shift that would result from a 1/4 wavelength displacement between the two antennas.

If the sources are collocated rather than displaced, the same effect is achieved by shifting the initial phase angle of the energy supplied to one antenna by $\pi/2$ radians; see figure 2-19. This might be accomplished by advancing the initial excitation of antenna 2 by 1/4 period.

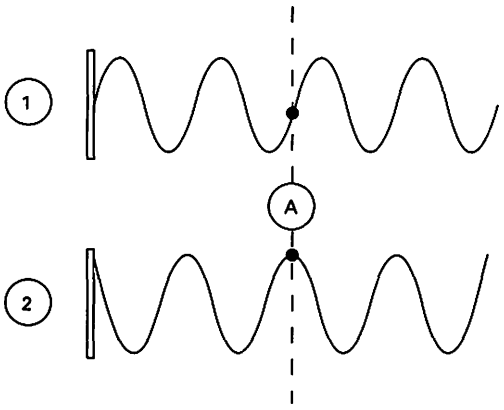


FIGURE 2-19 *Phase difference due to time difference*

In general, then, the form of equation 2-9 that accounts for some initial phase angle, ϕ, other than zero is:

$$E = E_0 \sin\left[\frac{2\pi}{\lambda}(r-ct)+\phi\right] \qquad (2\text{-}12)$$

Referring again to figures 2-18 and 2-19, it is now appropriate to discuss the total field strength at point A from the two sources. The total field strength is simply the mathematical summation of the instantaneous values of each individual field, $E_{1,2} = E_1 + E_2$. This summation can be expanded to any number of sources. If the net field strength is greater than that of the individual fields, the interference is constructive. Conversely, if the net field strength is less than that of the individual fields, the interference is destructive.

For example, if antennas 1 and 2 of figure 2-19 were exactly in phase, the net field strength would be twice the value as that from a single antenna (constructive interference). If the antennas had a phase difference of π radians (180°), the net field strength would be zero (destructive interference). Any other phase relationships will result in some intermediate net field strength.

Now consider a system in which two antennas are driven by a common source, except that the voltage reaching antenna 2 is advanced with respect to that reaching antenna 1, such that the radiation from antenna 2 is $\pi/2$ (90°) out of phase with that from antenna 1. How is the strength of the electric field at a point P (figure 2-20), determined when the point is equidistant from both antennas?

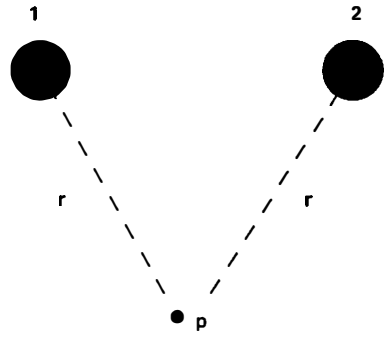


FIGURE 2-20 *Overhead view of two sources*

Since the antennas are both driven by the same source, radiation from each antenna will have the same wavelength and amplitude. If point P is located a distance r from both antennas, the electric field at point P due to the radiation from antenna 1 is represented as:

$$E_1 = E_0 \sin\left[\frac{2\pi}{\lambda}(r-ct)\right] \qquad (2\text{-}13)$$

The radiation reaching point P from antenna 2 is 90° out of phase with that from antenna 1; thus the electric field due to antenna 2 at point P is:

$$E_2 = E_0 \sin\left[\frac{2\pi}{\lambda}(r-ct)+90°\right] \qquad (2\text{-}14)$$

The total field strength at point P, E_P, due to the interference of these two waves is the algebraic sum of the two individual field strengths:

$$E_P = E_1 + E_2$$
or
$$= E_0 \sin\left[\frac{2\pi}{\lambda}(r-ct)\right] + E_0 \sin\left[\frac{2\pi}{\lambda}(r-ct)+90°\right] \qquad (2\text{-}15)$$

Apply the trigonometric identity:

$$\sin A + \sin B = 2\sin\frac{1}{2}(A+B)\cos\frac{1}{2}(A-B) \qquad (2\text{-}16)$$

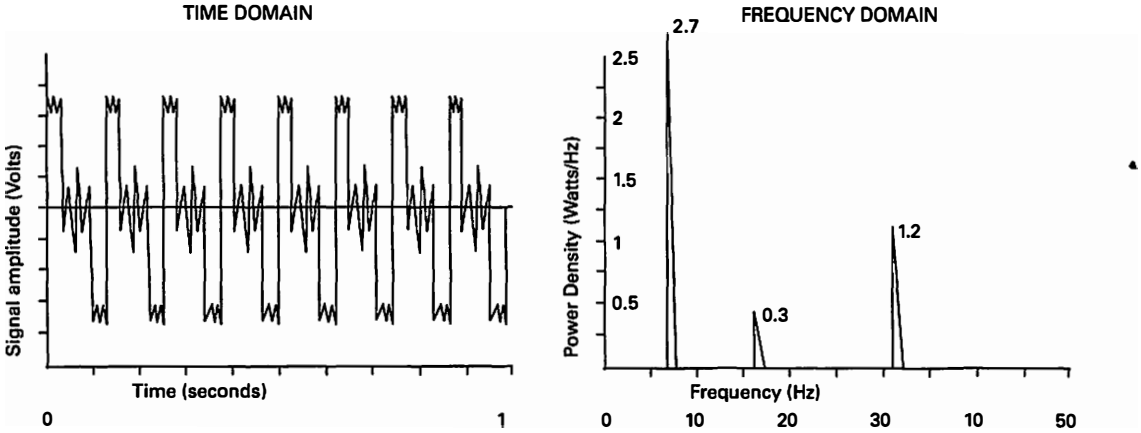


FIGURE 2-21 *A combination of multiple frequencies*

thus

$$E_P = 2E_0 \sin\frac{1}{2}\left[\frac{2\pi}{\lambda}(r-ct)+\frac{2\pi}{\lambda}(r-ct)+90^\circ\right] \times \cos\frac{1}{2}\left[\frac{2\pi}{\lambda}(r-ct)-\frac{2\pi}{\lambda}(r-ct)-90^\circ\right] \quad (2\text{-}17)$$

or

$$E_P = 2E_0\cos(-45^\circ)\sin\left[\frac{2\pi}{\lambda}(r-ct)+45^\circ\right] = 1.414E_0\sin\left[\frac{2\pi}{\lambda}(r-ct)+45^\circ\right] \quad (2\text{-}18)$$

Thus, in order to determine constructive or destructive interference, it is necessary to look at the absolute value of the amplitude of this resultant wave, E_P. In general terms this amplitude can be written:

$$E_P = \left|2E_0\cos\left(\frac{\phi}{2}\right)\right| \quad (2\text{-}19)$$

where:

ϕ = the phase difference between the radiating elements

(*Note*: Since cos is an even function, the negative sign is dropped from the phase angle.)

It is important to realize that the phase difference between any two waves can be the result of the phase difference of the source, the difference in path length traveled by the waves, or a combination of both. The phase difference caused by path length difference alone, Δx, is given by:

$$\phi = \frac{2\pi\Delta x}{\lambda} \quad (2\text{-}20)$$

By comparing this amplitude to what would have been received by a single source, we determine the type of interference by:

$$\left|2E_0\cos\left(\frac{\phi}{2}\right)\right| > E_0 \text{ (constructive)}$$

or

$$\left|2E_0\cos\left(\frac{\phi}{2}\right)\right| < E_0 \text{ (destructive)} \quad (2\text{-}21)$$

From the above inequalities it can be shown that constructive interference occurs when the phase difference is between 0° and 120°, or between 240° and 360°; and destructive interference occurs when the phase difference is between 120° and 240°.

Fourier Analysis (Principle)

Actual waves often have very complicated shapes and most periodic waves can be constructed from a combination of sine and cosine waves with different frequencies. In order to construct the wave, relative amounts of each frequency component must be known. In other words, the amounts of each frequency are like the ingredients on a recipe card for constructing waveforms. The reverse operation of finding the frequency components that make up a particular wave is called the spectral decomposition or frequency analysis.[v]

The real value of the frequency domain appears with signals that are complicated in the time domain, but consists of simple combinations of sinusoidal waves. Consider the time domain signal of figure 2-21. It is certainly not a smooth sine wave function, because it has multiple

frequency components. The frequency domain representation of this complicated time domain signal is given on the right in figure 2-21, and as shown, is simply a combination of three frequency components.

The strongest component is at 8 Hz, with additional contributions from 16 and 32 Hz. Because these higher frequencies are multiples of the original frequency, they are known as harmonics. The relative power levels are the square of the relative contributions to the amplitude. In this case, the 16 Hz contribution is 1/3 as large as the original at 8 Hz. Likewise, the 32 Hz contribution is 2/3 as large. Here's how these numbers were obtained:

ratio of 8 Hz to 32 Hz = 2.7:1.2 = 2.25:1
square root of 2.25 = 1.5:1 or 3:2

ratio of 8 Hz to 16 Hz = 2.7:0.3 = 9:1
square root of 9 = 3:1

The overall combination is therefore 3:1:2.[vi]

Losses

Losses in energy transmission are primarily due to spreading, absorption, and scattering. Spreading occurs as the wave travels outward from its source; its energy is expanding over an increasingly larger area. Energy per unit area of the wave front is then proportional to $1/R^2$, where R is the distance from the source.

If transmission is through some medium other than free space, the molecules of the medium will absorb some of the energy as the wave passes by. The severity of this attenuation (loss) due to absorption is dependent on the medium as well as the frequency of the radiation. Atmospheric attenuation is essentially negligible at the lower end of the radar frequency band, but at high radar frequencies it may become quite significant.

Scattering results from reflections off suspended particles within a medium, also causing attenuation of energy in the propagating wave. The terms *spreading, absorption,* and *scattering* apply to all types of waves traveling in a medium—electromagnetic, acoustic, and even explosive pressure-type waves.

Antennas

Maxwell's Theory

The essence of James C. Maxwell's theory is that an accelerating electric field will generate a time-varying magnetic field. Conversely, a time-varying magnetic field will generate a time-varying electric field. Thus, a changing electric field produces a changing magnetic field that in turn produces a changing electric field, and so on. This means that some form of energy transfer is taking place in which energy is transferred from an electric field to a magnetic field back to an electric field.

This energy transfer process also propagates through space as it occurs. Propagation occurs because the changing electric field creates a magnetic field that is not confined to precisely the same location in space, but extends beyond the limit of the electric field. The electric energy created by this magnetic field then extends somewhat farther into space than the magnetic field. The result is a traveling wave of electromagnetic energy.

Generation of Electromagnetic Waves

An elementary dipole antenna is formed by a linear arrangement of suitable conducting material. When the dipole is subjected to an alternating electric force, several phenomena occur that result in electromagnetic radiation.

An electric field is generated in the plane of a conductor that has a maximum intensity at the instant of maximum voltage. The associated lines of electric force are illustrated in the dipole schematic of figure 2-22a. Note that the outer lines are bowed away from the inner ones. This is due to a natural repulsion between lines of force in the same direction.

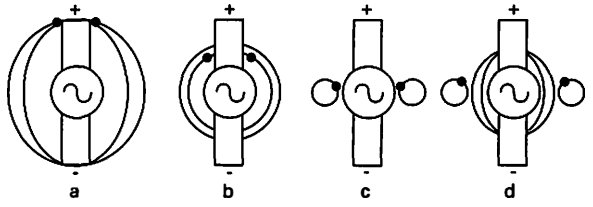


FIGURE 2-22 *Generation of an electric field about a dipole antenna*

In figure 2-22b the voltage has dropped somewhat, and the separated charges, together with their associated lines of force, have moved closer to the center of the dipole. As the voltage continues to drop, the lines begin to collapse back into the antenna.

The outermost lines do not collapse, however, and instead double upon themselves to form a closed loop. At this instant the voltage is zero, as shown in figure 2-22c. As the voltage begins to build in the opposite direction, the lines of force reform as shown in figure 2-22d. Interaction between these new lines and the closed loops results in the repulsion of the loops away from the antenna at the speed of light.

It was stated at the beginning of this section that an alternating electric field will produce a magnetic field. The antenna in the preceding discussion must therefore be surrounded by a magnetic field as well as by an electric one. A graphical illustration of this field is shown in figure 2-23. Since the current is maximum at the center of the dipole, the magnetic field is strongest at this point. The field is oriented in a plane at right angles to the plane of the electric field, as may be seen by comparing figures 2-22 and 2-23. Using the "right-hand rule," the magnetic field is as shown in figure 2-23.

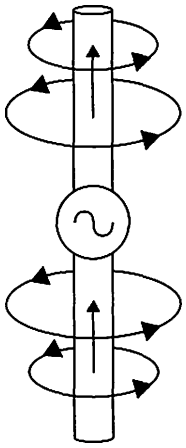


FIGURE 2-23 *Magnetic field about a dipole antenna*

The elementary dipole antenna of the foregoing discussion is the basic radiating element of electromagnetic energy. The constantly changing electric field about the dipole generates a changing magnetic field, which in turn generates a changing electric field, and so on. Because of the repulsion between lines of force in the same direction, the alternating fields thus produced are propagated out into space as electromagnetic waves.

The Electromagnetic Wave

In order to provide further insight into the nature of electromagnetic radiation, two approaches to the graphical representation of electromagnetic fields will be examined.

Figure 2-24 illustrates the waveform of both the electric (E) and magnetic (H) fields as they are mutually oriented in space. The frequency of these waves will be identical to the frequency of the voltage applied to the dipole. From previously developed relationships, their field strength at any distance and time may be defined as:

$$E = E_0 \sin\left[\frac{2\pi}{\lambda}(r - ct)\right] \qquad (2\text{-}22)$$

and

$$H = H_0 \sin\left[\frac{2\pi}{\lambda}(r - ct)\right] \qquad (2\text{-}23)$$

where:

E_0, H_0 = maximum field strength

c = speed of light = 3×10^8 meters/sec

Figure 2-25 illustrates the orientation of electromagnetic field components about a dipole antenna. The electric field (or flux) lines are represented by closed loops on either side of the antenna. Lines of magnetic flux are shown as opened and closed circles to indicate flux out of and

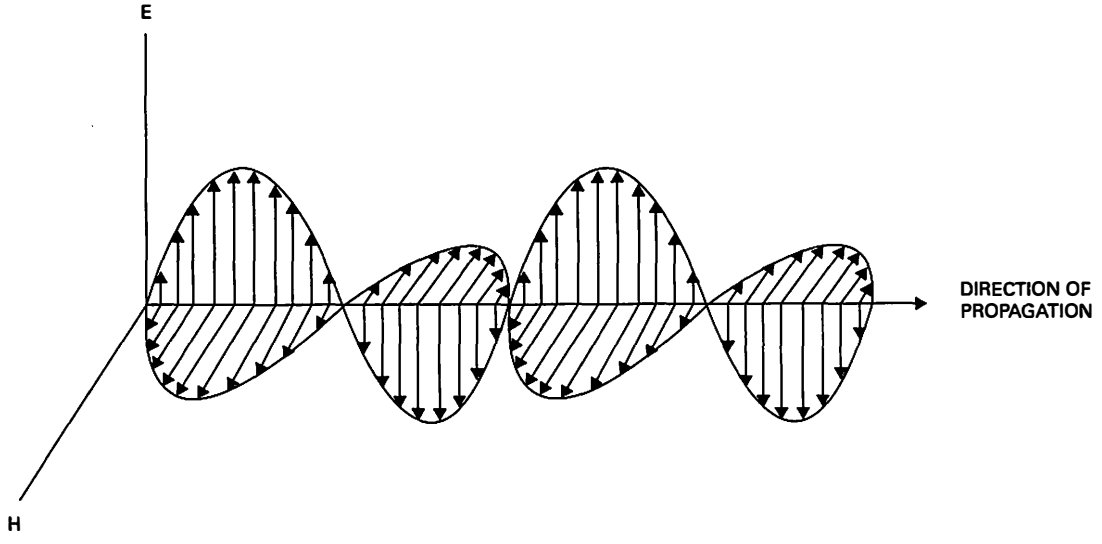


FIGURE 2-24 *Electromagnetic wave field orientation*

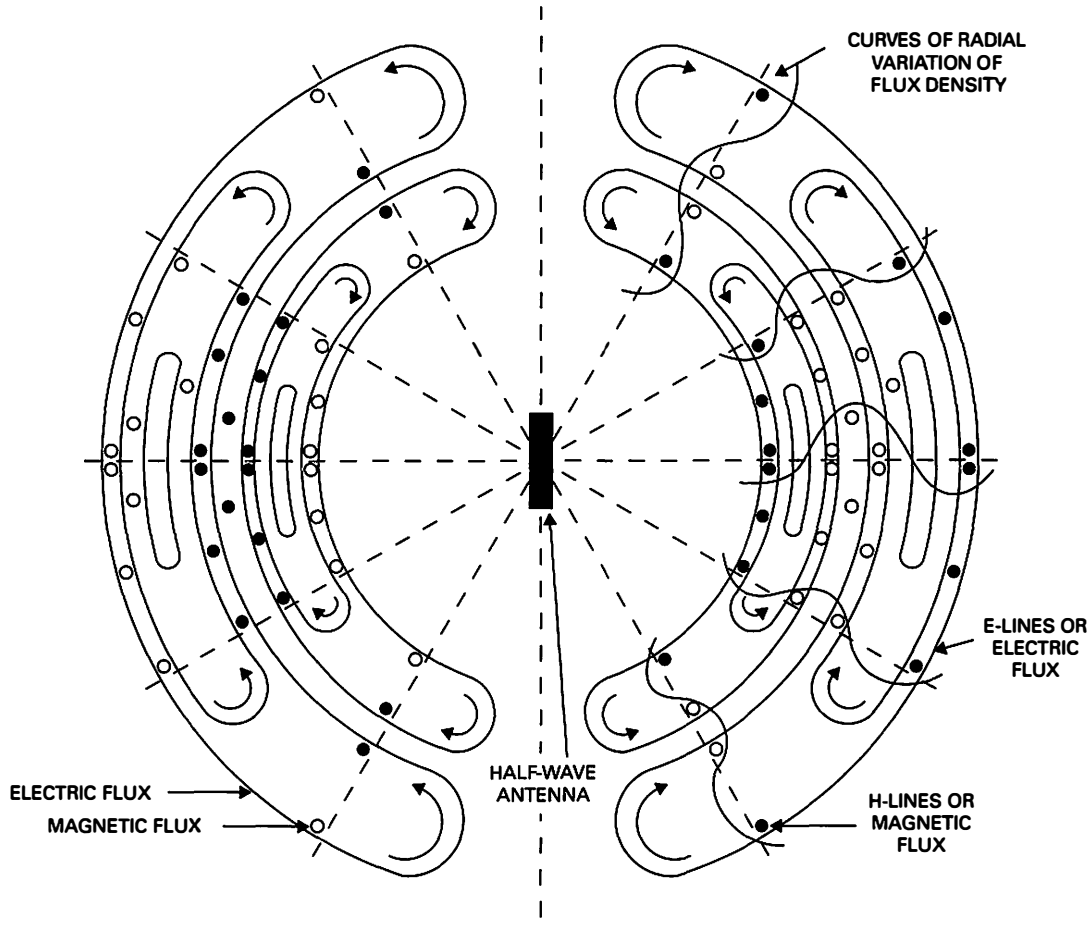


FIGURE 2-25 *Fields in space around a dipole antenna*

into the plane of the diagram. The darker waves labeled "curves of radial variation of flux density" indicate relative electric field strength at various radial directions from the antenna and would correspond to the *E* field curve of figure 2-24.

Note that both the *E* and *H* fields are strongest in the direction perpendicular to the antenna axis, as the amplitude of the "curves" shows. Conversely, there are only weak *E* and *H* fields propagated from either end of the antenna. Thus, antenna orientation is a major factor in electromagnetic energy transmission.

Wave Polarization

Energy from an isotropic antenna is radiated in the form of an expanding sphere. A small section of this sphere is called the wavefront, positioned perpendicular to the direction of the radiation field; see figure 2-26. Within this wavefront, all energy is in phase. Usually, all points on the wavefront are an equal distance from the antenna and the farther from the antenna the wave is, the less curved it appears. At a considerable distance the wavefront can be considered a plane surface at right angles to the direction of propagation.

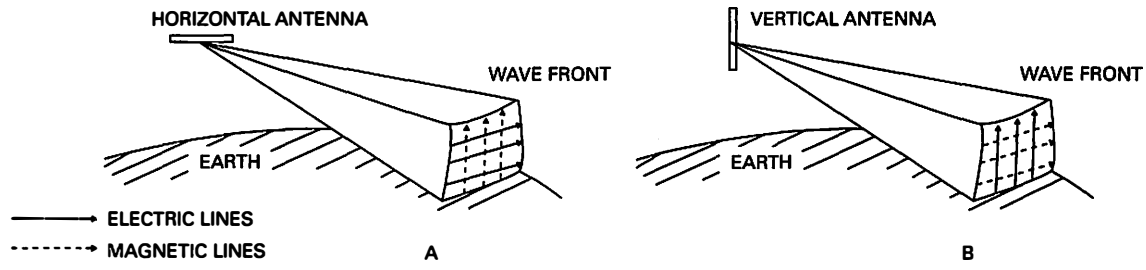


FIGURE 2-26 *Horizontal and vertical polarization*

As stated earlier, the radiation field is made up of magnetic and electric lines of force that are always at right angles to each other. By convention, the orientation of the electric field in the wave is called polarization. Most electromagnetic fields are said to be linearly polarized; however, some are circular. When the electric lines of force are horizontal, the wave is said to be horizontally polarized; and if the electric lines of force are vertical, the wave is said to be vertically polarized. Since the electric field is parallel to the axis of the dipoles, the antenna is in the plane of polarization.

In general, the polarization of a wave does not change over short distances. Therefore, transmitting and receiving antennas have the greatest sensitivity when they are oriented alike, especially if they are separated by short distances. When the transmitting antenna is close to the ground, it should be linearly polarized in the vertical direction, because vertically polarized waves produce greater signal strength along the earth's surface. On the other hand, when the transmitting antenna is high above the ground, it should be horizontally polarized to get the greatest signal strength possible to the earth's surface.[vii]

Over long distances the polarization changes. The change is usually small at low frequencies, but quite drastic at high frequencies. Additionally, for radar transmissions a received signal is actually a wave reflected from an object; therefore no set position of the receiving antenna is correct for all returning signals.

Linear polarization as described above is not the only possibility. Another type of polarization is called circular polarization. The best way to visualize circular polarization is helical or corkscrew shaped. The electric field rotates as it travels. If the rotation is clockwise in the direction of propagation, it is called right-hand circular polarization (RHCP); if counterclockwise, it is called left-hand circular polarization (LHCP). Antennas for circular polarization may look like corkscrews or pairs of dipoles at right angles; see figure 2-27. Circular polarization is often used in satellite communications because it eliminates the need to try and match the receiving antenna to the orientation of the satellite's antenna.



FIGURE 2-27 *Circular polarized antenna*

Some waves are not polarized at all. For example, sunlight is homogeneous mixtures of waves with all orientations and is said to be unpolarized. It can, however, become polarized either by filtering or upon reflection from a flat surface.[viii]

Antenna Beamforming

In almost all radar systems, it is desirable to concentrate the radiated electromagnetic energy into a directional beam, much akin to a flashlight beam. This makes possible the illumination of a specific area in space and subsequently allows for accurate determination of the location from which the reflected energy is returned. An additional advantage is the obvious power increase in the desired location.

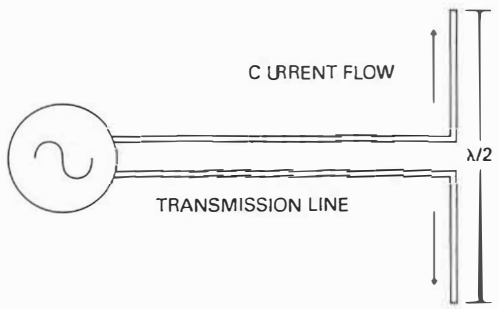


FIGURE 2-28 *Half-wave dipole antenna*

The simplest form of electromagnetic antenna is the half-wave dipole illustrated in figure 2-28. This antenna is created by bending the open ends of the transmission lines 90°. The physical length of this dipole antenna must be equal to one-half the wavelength ($\lambda/2$) for an ungrounded antenna of the transmission frequency in order to achieve optimum radiation. For a grounded antenna, the optimal length is one-fourth of the wavelength ($\lambda/4$) due to the reflection from the ground plane. Under these conditions, the radiation emitted from the antenna is in resonance with the exciting voltage.

The dipole antenna radiates its energy in all directions perpendicular to its axis. There can be no signal coming from the ends; therefore the electromagnetic energy field from a single dipole antenna resembles a huge doughnut

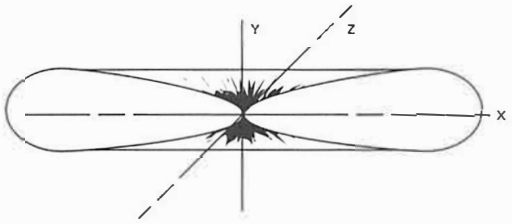


FIGURE 2-29 *Radiation cross section for a half-wave dipole antenna*

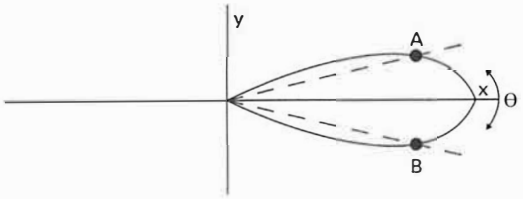


FIGURE 2-30 *Top view of the radiation pattern of a beam, looking at the X-Y plane. Points A and B represent where the half-power points are located.*

with the antenna at the center of the "hole"; see figure 2-29. In this sense, the dipole antenna has some directionality, or preferred direction. In cases where high sensitivity is required or when it is necessary to exclude transmission or reception from unnecessary directions, antennas can be made even more directional. The process of creating and controlling directionality in antennas is called beamforming. Beam forming has widespread applications in communications, radar, and sonar.[ix]

The radiation pattern for an ideal point source antenna is a sphere surrounding the antenna, with power distribution (in watts/meter^2) that is uniform, that is, the same everywhere. As seen from the radiation pattern of the dipole antenna in figure 2-29, the power is not equally distributed. It is maximum along the *X* axis and nulls out, or goes to minimum, along the *Y* axis. If observers were to walk around the *Y* axis of a fixed distance on the *X-Z* plane with a power meter, they would measure one constant value. If they stopped at any point and climbed a "ladder," that is, moved up parallel to the *Y* axis, they would see the power initially slowly decrease. The higher they went, the more rapidly the power would decrease until it reached almost zero somewhere near the 30°–40° angle off the horizontal. Thus, for any given distance away from the dipole, the power distribution in height (and depth) is not uniform.

What is needed is a way to concentrate the energy along the *X* axis and limit it to a small angle away from the axis. Thus, a beam is formed. The power density of the beam would not be uniform, but would change not only in height, as a dipole does, but also in width; see figure 2-30.

Thus, as observers walk around the *Y* axis, at a fixed distance from the axis, and measure the power received, they find a maximum power point along the *X* axis. Moving right or left shows the power level dropping off until it goes almost to zero. By convention, the width of the beam is measured as the angle between the points where the power has dropped to one-half the maximum (points *A* and *B*). These half-power points are routinely called the 3dB down points as well (3dB down comes from the fact that 10 log 1/2 equals –3). The beamwidth serves as a measure of the angular resolution of the radar.

Antenna systems therefore have two general purposes. They radiate radio frequency energy developed by the transmitter and confine this energy to a narrow beam. Subsequently, the antenna receives signals from a source and supplies them to the receiver. The antenna system must employ some type of a reflector or refractor for beam-forming radiated energy and for concentrating the received energy. How this is accomplished is by one of two methods: reflectors or through the use of arrays.

Quasi-Optical Reflectors

Earlier in the chapter, reflectivity of electromagnetic waves was discussed. This characteristic is put to use as a method for narrowing the beamwidth, thus increasing directivity. Various types and shapes of reflectors are used with driven elements and feeders to obtain increased directivity and a unidirectional radiation pattern. At microwave frequencies of radars, reflectors with parabolic cross sections are used primarily. These include the full paraboloid, parabolic cylinder, truncated paraboloid, and orange-peel paraboloid. These shapes are illustrated in figure 2-31. The geometry of a parabolic antenna and its associated feed system is illustrated in figure 2-32.

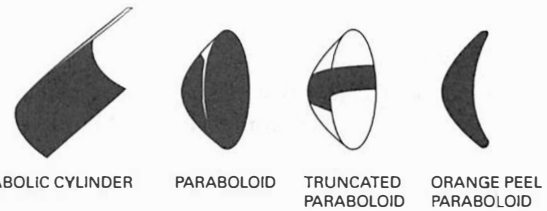


FIGURE 2-31 *Reflector shapes*

The paraboloid employs a point source of radiation, such as a feed horn or a dipole antenna with a small auxiliary backup reflector as a feeder, and produces a sharp

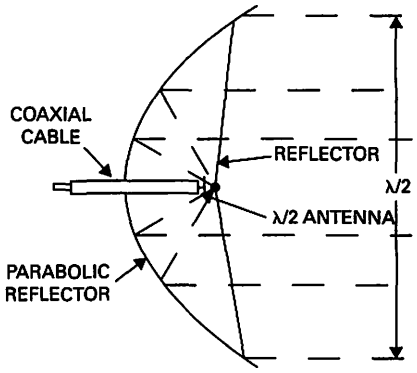


FIGURE 2-32 *Parabolic reflector*

parallel (pencil) beam. The parabolic cylinder uses a line source, such as a dipole array, and produces a flat, fan-shaped beam to cover greater volume of space for search purposes. In general, for a given frequency of radiation the larger the reflector is, the sharper the beam.

Reflectors are made of sheet metal or wire mesh. Figure 2-33 shows a radiation pattern for a parabolic cylinder. Comparative lengths of the major lobes and sidelobes are shown based on a maximum length of one. Angles shown are measured across the main beam.

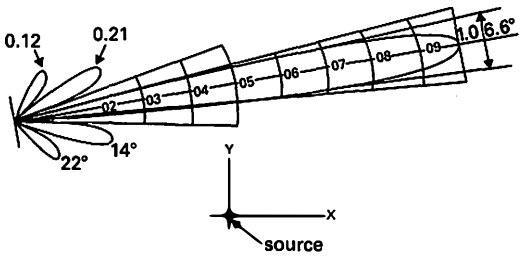


FIGURE 2-33 *Radiation pattern of a parabolic cylinder*

Linear Antenna Arrays

Another method of concentrating radar energy in order to obtain directivity is through the use of linear arrays. This method arranges two or more simple half-wave antennas in such a way that their fields add together in a specific direction and cancel in others. The principle of field summation and cancellation is the same as that discussed in the section in this chapter concerning interference. Such a set of elements constitutes what is referred to as a linear array, of which the two most common types are the broadside and the endfire arrays.

Consider the case of two half-wave dipole antennas spaced one-half wavelength apart and positioned as

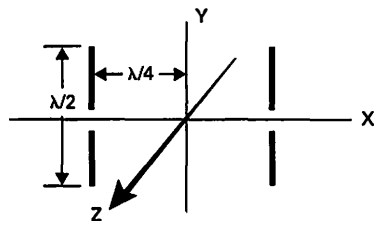


FIGURE 2-34 *Two-element linear array*

shown in figure 2-34. An investigation will be made of the radiation patterns from these antennas in the *X*, *Y*, and *Z* directions. The investigation is based upon excitation of both antennas by equal, in-phase voltages.

For clarity of presentation, the *Z* direction will be investigated first. The diagram of figure 2-35 has been rotated 90° about the *X* axis in order to present an end-on view of the two dipoles.

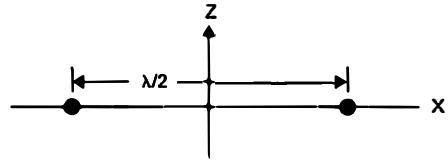


FIGURE 2-35 *End-on view of two-element array*

Since the dipoles are equidistant from the *Z* axis, the radiation from antennas *A* and *B* arriving at point 1 (or any other point on the *Z* axis) will be in phase. Each wave will have traveled an identical distance. The field amplitude at point 1 will therefore be a constructive combination of the fields of antennas *A* and *B* and will in fact be equal to twice the value resulting from either element alone. A similar analysis finds the same result true for the *Y* direction. Recall, however, that dipole antennas do not radiate significant amounts of energy along their linear axis and that the field strength will decrease correspondingly in the *Y* direction as the distance from the *Y* axis gets larger.

A significant feature of this linear antenna array becomes more obvious upon investigation of the *X* direction. Because the antennas are one-half wavelength apart, the waves from antenna *A* will arrive at a specific point on the *X* axis exactly 180° out of phase with the waves from antenna *B*. The resultant combination of the fields will produce a net field strength of zero.

It should now become evident that the proper spacing of multiple dipole antennas accomplishes two distinct means of directing electromagnetic energy in a desired pattern and direction. The first of these is reinforcement of field strength in the desired direction, and the second is the cancellation or reduction of field strength in directions

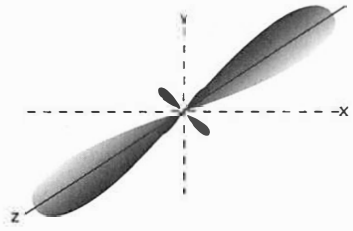


FIGURE 2-36 *Radiation pattern for a two-element linear array*

where no field is desired. The resulting pattern of electromagnetic radiation from the two-antenna system is shown in figure 2-36.

Note that there are main lobes in both directions along the Z axis as well as sidelobes.

The main lobe is the carrier through which the radar unit sends and receives information. It is therefore desirable to have most of the available energy concentrated into this lobe. Furthermore, since bearing information provided by the radar is determined by the direction in which the main lobe is pointing, a radar return from a sidelobe would provide erroneous and thus undesirable information. In order to reduce the possibility that this situation will occur, the sidelobes are reduced in strength as much as possible.

Broadside array The design of a broadside antenna array such as the 16-element (dipole) array in figure 2-37 is based upon the principles discussed in the preceding section. The directivity pattern is along the $+Z$ and $-Z$ direction (perpendicular to the page). Since the antennas are excited in phase, the constructive interference occurs in the direction of $\pm Z$, whereas destructive interference is established along $\pm Y$ and $\pm X$.

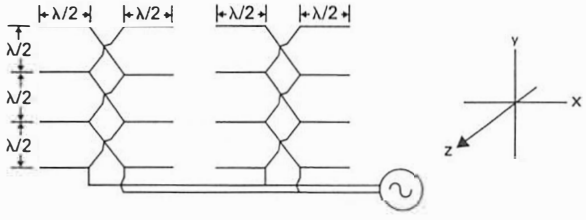


FIGURE 2-37 *Broadside array*

The radiation pattern of this antenna is similar to that of the double half-wave dipole discussed earlier. Field strength of this pattern along the $\pm Z$ direction will, of course, be greater than the two-dipole case due to the increased number of radiating elements.

Endfire array To this point the discussion of multielement arrays has dealt only with antenna elements that are stimulated with the same input at the same time—that is, in phase stimulation. Consider what would happen to the broadside array of figure 2-37 if adjacent elements were stimulated 180° out of phase. The constructive interference would now occur along the $\pm Y$ axis instead of the $\pm Z$ axis. The directivity of this antenna will be in the plane of this text and toward the top and the bottom of the page.

This discussion of phase shift in antenna excitation leads to a consideration of electronic beam positioning. By shifting the excitation of adjacent elements 180° out of phase, the main beam was rotated 90° in bearing from the $\pm Z$ direction to the $\pm Y$ direction. It is logical to assume that some intermediate beam directivity can be readily achieved through a suitable phase shift in element excitation. This principle is, in fact, used and will be discussed in greater detail in later chapters.

Parasitic elements The directivity patterns of both broadside and endfire type arrays are still not optimum, since the main beams go in two directions. It is obviously desirable that the main radar beam be concentrated in one direction only. One method of accomplishing this is through the use of parasitic elements. A parasitic element is simply a conductor of suitable length that is placed in the field near the radiating antenna (driven element) and aligned in the same direction. The field of the radiating (or driven) antenna causes a voltage to be induced in the parasitic element. This voltage will produce a current flow and, if the length of the parasitic element is close to one-half wavelength, the element will generate some electromagnetic radiation of its own.

The phase of the voltage induced in the parasitic element is such that there will be some degree of cancellation on the side of the radiating antenna and reinforcement on the opposite side. Figure 2-38 illustrates the radiation patterns from a half-wave antenna alone and from an antenna-and-parasitic-reflector combination. Essentially, the parasitic element serves as a reflector and will modify the radiation pattern accordingly in a specific direction.

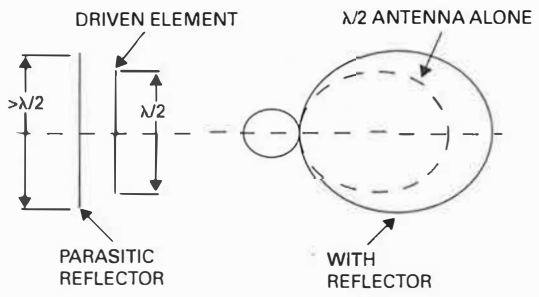


FIGURE 2-38 *Radiation patterns with and without a parasitic reflector*

Electromagnetic Wave Modes of Propagation in the Atmosphere

Electromagnetic waves that travel in the earth's atmosphere are subject to many factors that affect the range of the signal. Electromagnetic waves suffer from reflection from the ground, mountains, and other large objects like buildings. Because of the numerous layers in the atmosphere, all having different indexes of refraction, these waves also bend, thus affecting the range of the signal. In addition, tall objects will cause diffraction. (How else is reception behind a building or a mountain explained?) The phenomena of reflection, refraction, and diffraction are all frequency dependent and therefore each affects how a wave will propagate. Table 2-2 shows the major breakdown of communication bands and their carrier frequency range.

TABLE 2-2 *Communication bands*

Frequency Range	Band Designation	Propagation Mode
30–3000 Hz	ELF	Ground waves
3–30 kHz	VLF	
30–300 kHz	LF	
300–3000 kHz	MF	Sky waves
3–30 MHz	HF	
30–300 MHz	VHF	Space waves
300–3000 MHz	UHF	
3–30 GHz	SHF	
30–300 GHz	EHF	

Ground Waves

At lower frequencies and with vertical polarization, an electromagnetic wave will travel along the earth's surface; see figure 2-39. The front of a ground wave is connected to the surface of the earth at one end and to the ionosphere at the other. The ionosphere is the region above the troposphere (where air is), from about 50 to 250 miles above the earth.

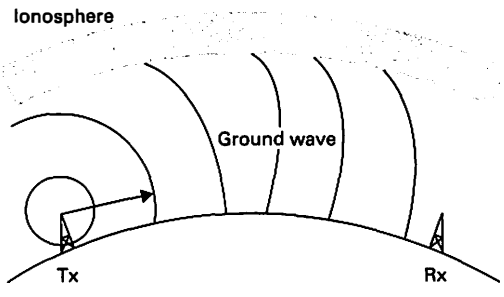


FIGURE 2-39 *Ground wave propagation*

The ionosphere is a collection of ions, which are atoms that have some of their electrons stripped off, leaving two or more electrically charged objects. The sun's rays cause these ions to form, after which they slowly recombine. The propagation of radio waves in the presence of ions is drastically different from their propagation in air. With a highly conductive surface such as that provided by salt water and ions in the ionosphere, attenuation will be minimized and ranges will be extended more than that experienced over dry ground. Ground waves travel between these two limits, acting like a duct.[x]

Very-long-range propagation is possible using ground waves. At frequencies in excess of 5 to 10 MHz, excessive bending due to diffraction results in the collapse of the electrical field and complete attenuation occurs. Ground wave propagation requires wavelengths that are too long for anything but communications by ships and aircraft, but can be used by fixed, ground-based radar.[xi]

Sky Waves

Radio waves in the Low Frequency (LF) and Medium Frequency (MF) ranges may propagate as ground waves, but suffer significant losses, or may be attenuated, particularly at higher frequencies. As the ground wave mode fades out, a new mode develops, the sky wave. Layers in the ionosphere of altitudes from 50 to 500 kilometer can be used to refract electromagnetic energy back toward the surface of the earth where it can be reflected to the layer again; see figure 2-40. Using a series of these wave reflections and refractions, communications can be established at long ranges, up to hundreds of miles, but not with complete reliability.[xii]

Sky waves in the LF and MF frequency bands are usually possible only at night, when the concentration of ions is not too great, since the ionosphere also tends to attenuate the signal. However, at night, there are just enough ions to reflect the wave but not reduce its power too much.[xiii]

The High Frequency (HF) band operates almost exclusively with sky waves. The higher frequencies have less attenuation and less refraction in the ionosphere as compared to MF. At the high end, the waves completely penetrate the ionosphere and become space waves. At the low end, they are always reflected. The HF band operates with both these effects almost all of the time. The characteristic of sky wave propagation depends on the conditions in the ionosphere, which in turn are dependent on the activity of the sun.[xiv]

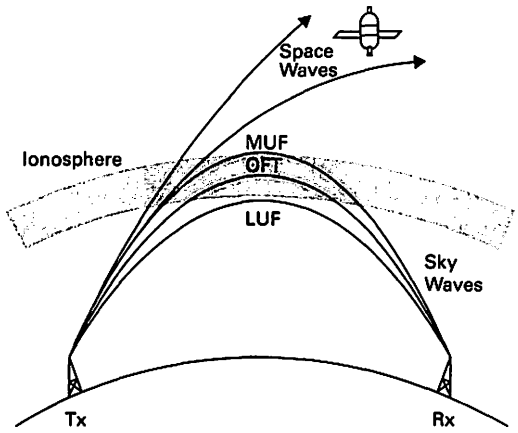


FIGURE 2-40 *Sky wave propagation*

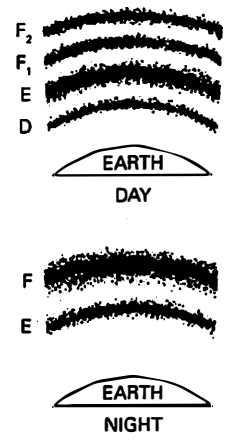


FIGURE 2-41 *Regions of the ionosphere*

Care must be exercised in selecting the layer to be used and the frequency and the elevation angle of transmission, to ensure that refraction occurs in the layer selected and the waveform returns to earth at the desired point. Again, the frequencies involved are too low to be used for anything but communication by ships or aircraft or ground stations for over-the-horizon target detection.

The ionosphere has several well-defined regions that are used for sky wave communications, as shown in figure 2-41.

- *D region:* about 75–95 km. Relatively weak ionization. Reflects Very Low Frequency (VLF) waves for long-range communications; refracts LF and MF for short-range communications; has little effect on Very High Frequency (VHF) and above; gone at night.
- *E region:* 95–150 km. Depends upon angle of the sun: refracts HF waves during the day up to 20 MHz to distances of 1,200 miles; greatly reduced at night. A strong effect of ionospheric scattering of VHF communications.
- *F region:* 150–400 km. Has separate F_1 and F_2 layers during the day. F_1 layer density depends on the angle of the sun; its main effect is to absorb HF waves. F_2 layer provides long-range HF communications; very variable; height and density change with time of day, season, and sunspot activity. Responsible for reflection of HF radio waves.[xv]

Since the propagation characteristics depend on frequency, several key frequencies are defined.

Critical frequency The critical frequency is the minimum frequency that will penetrate the ionosphere at vertical incidence. The critical frequency increases during the day and decreases at night. At other angles, the wave will be refracted back. At frequencies above the critical frequency, some range of frequency with vertical incidence and down will become space waves. This will cause a gap in coverage on the ground known as a skip zone.

Maximum usable frequency (MUF) The higher the frequency of a radio wave, the lower the rate of refraction by the ionosphere. Therefore, for a given angle of incidence and time of day, there is a maximum frequency that can be used for communications between two given locations; see figure 2-40.

Lowest usable frequency (LUF) As the frequency of a radio wave is lowered, the rate of refraction increases. So a wave whose frequency is below the established LUF is refracted back to earth at a shorter distance than desired, as shown in figure 2-40. Atmospheric noise is also greater at lower frequencies, reducing the signal-to-noise ratio.

Optimum working frequency (OWF) For two stations, the most practical operating frequency is one that can be relied on to have the least number of problems. It should be high enough to avoid the problems of fading and noise encountered at the lower frequencies, but not so high as to be affected by the adverse effects of rapid changes in the ionosphere. A frequency that meets the above criteria is known as the optimum working frequency.

Space Waves

Above 30 MHz the ionosphere will not refract electromagnetic waves back toward earth, and there is essentially no ground wave propagation. In a vacuum, energy at

these frequencies would travel in a straight line. However, due to changes in atmospheric density, electromagnetic waves are refracted toward the earth, but not sufficiently to follow the earth's curvature. Nevertheless, the frequency is still low enough for some significant effects that affect propagation ranges.

Ionospheric scatter is a propagation effect caused by the E region of the ionosphere scattering a signal in all directions; see figure 2-42. The energy that is scattered back toward the earth's surface may be usable and can be received effectively at ranges up to 600–1000 miles.

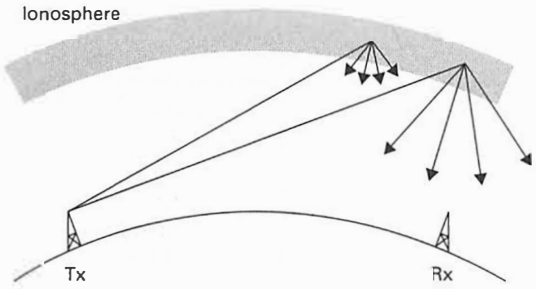


FIGURE 2-42 *Ionospheric scatter propagation*

Tropospheric scatter is caused by particles in the troposphere that, like ionospheric scatter, direct signals in multiple directions, some being reflected back toward the earth's surface and that can be received up to ranges of 400 miles see figure 2-43. An example of this effect is vehicle headlights in the fog.

Tropospheric ducting causes unusually long ranges for electromagnetic energy transmissions to be observed at sea. This ducting situation arises because of abnormal atmospheric conditions a few miles above the earth. Normally, the warmest air is found near the surface of the water. The air gradually becomes cooler as the altitude increases.

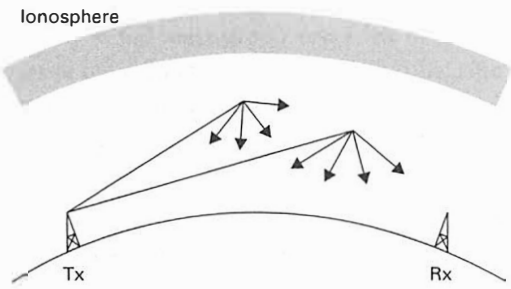


FIGURE 2-43 *Tropospheric scatter propagation*

Sometimes, unusual situations develop where warm layers of air are found above cooler layers. This condition is known as a temperature inversion. Because electromagnetic waves travel slower in cold, dense air than in warm air, the wave is naturally bent back to the ground. When the refraction from the inversion matches the curvature of the earth, long ranges can be achieved. The wave acts as though it is trapped in a duct just above the surface of the earth. This ducting always occurs to some extent and improves the range over the true line-of-sight by about 10%.

With certain exceptions, ducts are formed over water where the following conditions are observed aboard ship:

1. A wind is blowing from the land.
2. There is a stratum of quiet air.
3. High pressure, clear skies, and little wind is present.
4. Cool breeze exists over warm open ocean.
5. Smoke, haze, or dust fails to rise.
6. Received signal is fading rapidly.

Radar Line-of-Sight

At the high-frequency ranges normally used for radar transmission, space wave propagation is normal. Transmission and reception are therefore limited to "line-of-sight" with some bending due to refraction and the other effects of space wave transmissions.

At the earth's surface the "radar line-of-sight horizon" is the maximum allowable separation between transmitter and receiver and can be readily calculated by the following empirical relationship:

$$R(\text{km}) = \sqrt{17H_T} + \sqrt{17H_R} \qquad (2\text{-}24)$$

where R is the maximum separation in kilometers and H_T and H_R are target and receiver heights in meters, respectively. The above equation takes into account the bending of the ray due to refraction see figure 2-44. Analysis of the equation shows an approximate distance increase of 16% over the direct or geometric line-of-sight that is given by the following relationship:

$$\text{geometric } R(\text{km}) = \sqrt{13H_T} + \sqrt{13H_R} \qquad (2\text{-}25)$$



FIGURE 2-44 *Extension of radar horizon as the result of atmospheric refraction of space waves*

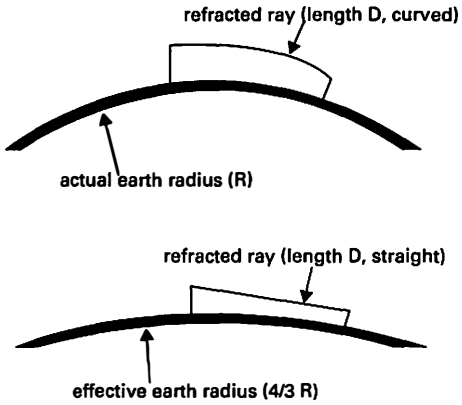


FIGURE 2-45 *Effective earth radius*

Solving the Radar Line-of-Sight Equation

Why is it that when you input antenna height in meters, the answer turns out to be in kilometers?

The short answer is that the conversion factor is already in the equation; the longer answer involves solving two different problems. The first is to account for the refraction of the radio waves across the curvature of the earth. The second problem is simple geometry.

To account for refraction, measure a refracted ray as it curves over the earth; this is length D over the curved distance between two points, see figures 2-45 and 2-46. Next, straighten length D into a straight line, and calculate the radius of a circle required to connect the same two points without refraction; this gives a new radius, R_e. R_e is the effective earth radius without refraction and 4/3 the earth's actual radius. The approximate radius of the earth is 6,375 kilometers; therefore R_e is equal to 8,500 kilometers.

The second step is a geometry problem involving a triangle and finding the distance $d = d_1 + d_2$ by use of the

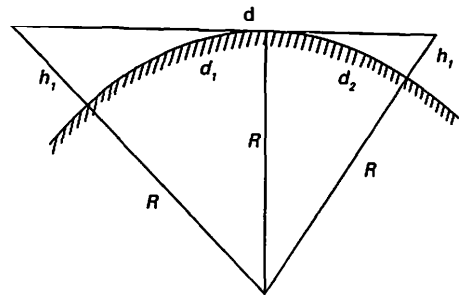


FIGURE 2-46 *Radar line-of-sight geometry*

Pythagorean Theorem; d_1 and d_2 are given by:

$$(d_1)^2 = (R_e + h_t)^2 - R_e^2 = h_t^2 + 2R_e h_t$$
$$(d^2)^2 = (R_e + h_r)^2 - R_e^2 = h_r^2 + 2R_e h_r$$

The height of the transmitter or receiver squared is small compared to the effective earth radius (h_t and $h_r << R_e$); therefore the equation can be reduced to:

$$(d_1)^2 = 2R_e h_t \quad \text{and} \quad (d_2)^2 = 2R_e h_r$$
$$d = \sqrt{2R_e h_t} + \sqrt{2R_e h_r}$$

where:

$R_e = 8{,}500{,}000$ meters

$$d = \sqrt{2(8{,}500{,}000\ \text{m})(h_t)} + \sqrt{2(8{,}500{,}000\ \text{m})(h_r)}$$
$$d = \sqrt{17{,}000{,}000\ \text{m}\,(h_t)} + \sqrt{17{,}000{,}000\ \text{m}\,(h_r)}$$
$$d = \sqrt{17h_t}\,(1{,}000\ \text{m}) + \sqrt{17h_r}\,(1{,}000\ \text{m})$$
$$d = \sqrt{17h_t}\ \text{km} + \sqrt{17h_r}\ \text{km}$$

---

[i] Hall, 16
[ii] Hall, 14
[iii] Hall, 17
[iv] Hall, 27
[v] Hall, 17
[vi] Hall, 19
[vii] *Electronics Technician Volume 7*, 2–2
[viii] Hall, 29–30
[ix] Hall, 30
[x] Hall, 33
[xi] Hall, 34
[xii] Hall, 34
[xiii] Hall, 34
[xiv] Hall, 34
[xv] *Electronics Technician Volume 7*, 1–10

3 Radar Fundamentals

Introduction

Radar is an acronym for *radio detection and ranging.* The development of radar started in 1886 when Heinrich Hertz demonstrated that radio waves could be reflected by metallic objects. In 1904 a German engineer named C. Hulsmeyer patented an "obstacle detector and ship navigation device." In 1922 U.S. Naval Research Laboratory (NRL) engineers A. H. Taylor and L. C. Young detected a wooden ship using continuous-wave radar. The first detection of an aircraft came in June 1930 by L. A. Hyland, also of NRL. In England, Sir Robert Watson-Watt was asked by the Royal Academy of Britain in 1932 if a radio ray could be directed as a weapon against enemy aircraft. He answered the question by saying, in effect, "No, but we certainly can locate the aircraft."

Through the use of electromagnetic waves, certain objects can be detected and located at far greater distances than possible by the unaided human eye. This detection is unimpaired by night, fog, smoke, clouds, dust, or most other obstacles to ordinary vision. Moreover, radar permits a rapid, convenient, and accurate measurement of range to the object that it "sees" in a way unequaled by most optical devices and methods. Radar can also measure the relative velocity of objects in a simple and natural way. Range and range rate (change in range per unit time) information can then be fed directly into fire control systems with precision unobtainable by a human operator.

Other aspects of radar performance are definitely inferior to that of the eye. The detailed definition of a scene that radar offers is much poorer than that afforded by the eye. Most radar equipment can show only the gross outlines of objects such as ships, vehicles, or aircraft. Two closely spaced objects such as aircraft may also appear as one object to the radar. Radar is therefore best suited to dealing with isolated targets that appear against a relatively featureless background, such as ships at sea, airplanes in the sky, islands, or any large terrain features.

Radar systems employ two basic types of energy transmissions: pulse and continuous wave (CW). Pulsed radar transmits radio frequency energy in a series of short pulses separated by nontransmission intervals or rest times. Target returns are processed during these nontransmission intervals, and range is determined based on the total travel time for the pulse-return cycle.

In CW radar, on the other hand, the transmitter sends out a continuous signal. If a nonmoving object is in the path of the transmitted wave train, the frequency of the reflected signal will be the same as the transmitted signal. If the object is moving, the frequency of the reflected signal will differ from the transmitted signal, and that difference can be used as an indicator of target motion. Consequently, with basic CW transmission, either a movement of the radar or the target is necessary to produce an indication of target presence.

Principles of Operation

Pulse Transmission

The pulse-echo system of detection is used in almost all radar sets. In this system, the transmitter is turned on for short periods and off for long periods of time. During the period when the transmitter is turned on, it transmits a short burst (pulse) of energy. When a pulse strikes an object, part of the reflected energy returns to the receiver, where it is processed and displayed. Since the transmitter is turned off after each pulse, it does not interfere with the receiver, as it might if a continuous signal were used.

The elapsed time between the beginning of one pulse transmission and the beginning of the next is called the pulse repetition time (PRT) as shown in figure 3-1.

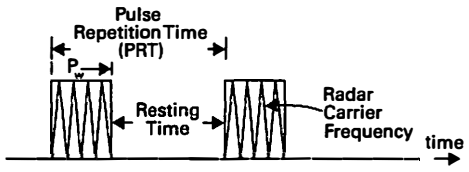


FIGURE 3-1 *Pulse transmission*

This time must be of sufficient duration to allow the echo pulse to return from the maximum range of the system; otherwise reception of a return will be obscured by the succeeding pulse. An expression commonly used instead of PRT is that of the pulse repetition frequency (PRF), expressed in hertz, where PRF = 1/PRT.

The minimum range at which a target can be detected is determined by the pulse width, PW. If a target is very close to the transmitter such that an echo is received before the transmitter is cut off and the receiver turned on, the echo will not be displayed. Because the pulse duration time must be short to increase reception of nearby targets and yet contain sufficient power to ensure a return echo of enough magnitude from the maximum range of the system, extremely large transmitted power outputs are required to produce a pulse of sufficient energy. The useful power of the transmitter is contained in the radiated pulses and is termed the peak power of the system.

The following definitions and abbreviations are often used with a pulsed radar type of system:

- Radio frequency (RF)/carrier frequency (f)—the working/operating frequency. The frequency of the carrier wave being modulated forming the pulse train. Units: Hz (GHz or MHz range)
- Pulse repetition frequency (PRF)—number of pulses transmitted per second. Units: Hz
- Pulse width (PW)—the active transmit time or duration of the EM (electromagnetic) pulse. Units: μsec
- Rest time (RT)—the nontransmit time or the interval between EM pulses. Units: μsec
- Pulse repetition time (PRT)—the total time for one transmission cycle. PRT = PW + RT or PRT = 1/PRF. Units: μsec

Since the radar system is not transmitting for a long period during its total cycle, the average power is very low compared to the peak power during the pulse time. The relationship between average power dissipated over the entire cycle and peak power developed during the pulse can be expressed by the following equation:

$$\frac{\text{average power}}{\text{peak power}} = \frac{\text{PW}}{\text{PRT}} \qquad (3\text{-}1)$$

and, since PRF = 1/PRT, the equation is alternately expressed as:

$$\frac{\text{average power}}{\text{peak power}} = (\text{PW})(\text{PRF}) \qquad (3\text{-}2)$$

The relationship between pulse width and pulse repetition time and between average power and peak power is termed the duty cycle (DC). It represents the ratio of the time that the transmitter is on to one complete cycle of operation of the transmitter.

$$\frac{\text{PW}}{\text{PRT}} = \frac{\text{average power}}{\text{peak power}} = \text{duty cycle} \qquad (3\text{-}3)$$

EXAMPLE 3-1: What is the duty cycle for a 2 μsec pulse repeated at a PRF of 5,000 Hz, and what is the peak power of this system if the average power is 20 kilowatts?

SOLUTION:

$$\frac{1}{\text{PRF}} = \frac{1}{5{,}000}\ \text{sec} \quad \text{or} \quad 200\,\mu\text{sec}$$

$$\frac{\text{PW}}{\text{PRT}} = \frac{2}{200} = 0.01 = \text{duty cycle}$$

$$\text{peak power} = \frac{\text{average power}}{\text{duty cycle}}$$

$$= \frac{20\ \text{kilowatts}}{0.01} = 2{,}000\ \text{kilowatts}$$

The relationship is graphically presented in figure 3-2.

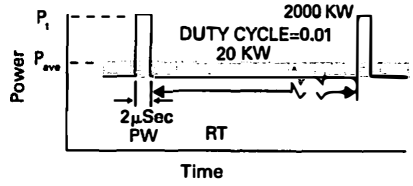


FIGURE 3-2 *Transmitted power distribution*

Range Determination

The effectiveness of range determination depends primarily on the ability of the system to measure distance in terms of time. Electromagnetic radiation travels in space at a constant velocity of about 3×10^8 m/sec. When it is reflected, there is no loss in velocity, merely a redirecting of the energy path.

Range is then determined by the time required for a two-way energy transmission. The determination of range is made by using the equation:

$$R = \frac{ct}{2} \qquad (3\text{-}4)$$

where:

R = distance of range from radar set to target
c = propagation speed of electromagnetic waves
t = time required for the two-way trip

Since range determinations are based on the measurement of time required for a pulse to travel to a target and return, radio waves traveling at 3×10^8 m/sec, will, for example, travel 300 meters in a microsecond. This corresponds to a time of 3.3 microseconds to travel 1 kilometer. If the aircraft shown in figure 3-3 exists at a distance of 1 kilometer from the radar system, the reflected echo will reach the receiver 6.7 microseconds after transmission. This interval is the round-trip time and can be converted to linear range measurements using equation 3-4.

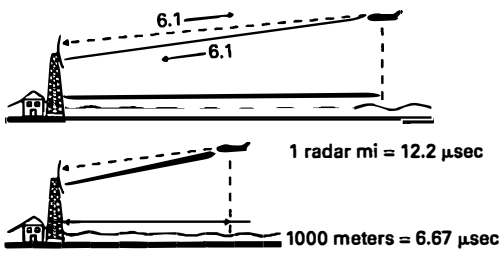


FIGURE 3-3 *Radar range-time relationship*

Mechanization

Radar systems vary considerably in detail as well as in the function they are designed to accomplish. However, the basic principle of operation is essentially the same for all systems, and a single radar can be visualized in which the functional requirements hold equally well for all types of radars.

Typical pulse-echo-type radar is composed of seven basic components—transmitter, receiver, power supply, synchronizer, duplexer, antenna system, and display. The relationship among these components is illustrated in figure 3-4.

Transmitter

The transmitter's function is to generate high-power pulses under the control of a timer. The carrier wave is produced by a radio frequency oscillator, which in most radar systems is called a magnetron. Since a magnetron produces a continuous waveform, it is necessary for a modulator or pulse generator to modulate the output. The term modulator refers to circuits that control the application of high voltage to the magnetron to create pulses at a predetermined time and duration.

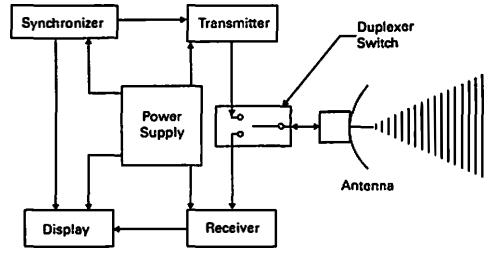


FIGURE 3-4 *Components of a basic radar system*

Receiver

The receiver converts incoming electromagnetic waves at a tuned frequency into electrical signals that can be used by a signal processor. The frequency and bandwidth characteristics of the receiver are matched to the transmitted pulses to maximize the detection of the return signal.[i] The signal processor contained in the receiver attempts to discriminate the target from the noise. The more sensitive the receiver is without simultaneously introducing more noise, the greater the range of detection.

Power Supply

The power supply furnishes all the electrical (AC and DC) voltages necessary for the operation and interaction of the system components. Pulse power generated by the transmitter may be much greater than the average power supplied because energy can be stored during the rest time between pulses in devices such as capacitors.

Synchronizer

The synchronizer, also known as the timer, supplies the controlling signals that determine the timing of the transmitted pulses and coordinates the action among all circuits in the system. This device establishes the pulse repetition frequency of the radar system, ensures that the modulator and display indicator group operate in a definite time relationship with each other, and initiates their functions at the instant that the transmitter produces a pulse of RF energy.

Duplexer

Most radar sets employ a single antenna for both transmission and reception. The duplexer, commonly known as the transmitting/receiving (TR) switch, is a device that allows the antenna to be sequentially connected to the transmitter and then the receiver. Its main function is to prevent high transmitting power from entering and thus damaging the receiver.

Antenna

The antenna system receives the radio frequency energy from the transmitter and radiates it in the form of a highly directional beam. Any returning echoes are received by the antenna and passed on to the receiver. Movable antennas are mounted on platforms positioned by syncrotransmitters that measure the antenna orientation in space. Stationary antennas are able to form beams and establish target orientation by signal-processing techniques as explained in the electronic scanning section in chapter 4.

Display

The display produces visual indications of the received pulses to furnish information to operators. The type of information usually desired includes range, azimuth or bearing, and the elevation of returned echoes. As stated earlier, range to a target is determined from the measurement of time from transmission to reception. Bearing and elevation of the target is a function of antenna bearing and elevation angle.

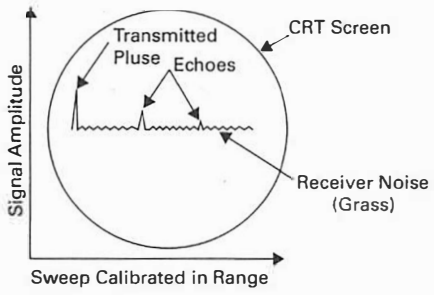


FIGURE 3-5 *A-scan display*

The simplest display type is called the A-scan. This type of display has range or time on the horizontal axis and the amplitude of the return signal on the vertical axis; see figure 3-5. The bearing and range display is the most common display, commonly known as the plan position indicator (PPI). This display presents information from a "bird's-eye view" looking down from above and showing range from the center of the display outward and

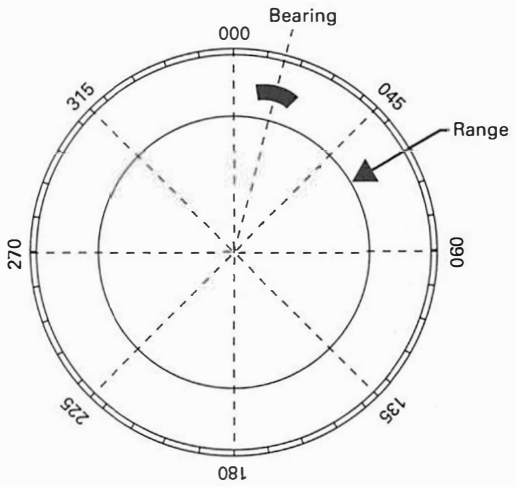


FIGURE 3-6 *PPI display*

bearing, either relative or true in the clockwise direction; see figure 3-6.

Note: The terms *bearing* and *azimuth* refer to angular distances in the horizontal plane and are often misused. Bearing is the generally used term and may be used to indicate angular distance of a target from a ship's heading (relative bearing) or from true north (true bearing). Azimuth is the true bearing of a point—that is, the bearing from a true north line.

Wave Guide

At extremely high frequencies such as those employed in all but the lowest frequency radars, wire conductors and even coaxial lines can act like a short circuit across the output of the transmitter. Under these conditions, a simple hollow metal tube or wave guide containing an inert dielectric gas, such as nitrogen or dry air, provides the most efficient means of connecting the radar transmitter to the antenna; see figure 3-7.

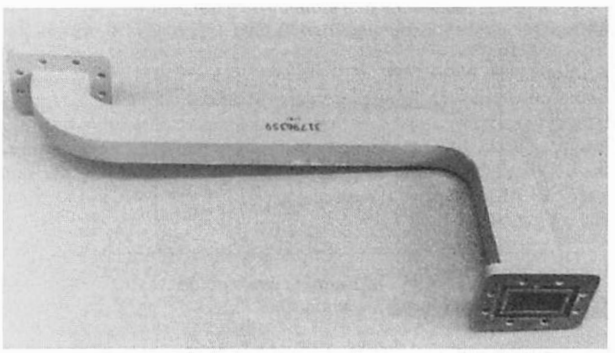


FIGURE 3-7 *Wave guide*

The wave guide can be of rectangular, round, or oval in cross section, with the major dimension approximately equal to half the wavelength at the transmitted frequency. The minor dimension is chosen based on desired attenuation loss and power handling capacity.

In the event that a frequency much lower than designed is employed, cutoff will occur, and transmission will cease. At high power levels, loss of the dielectric constant or insulating ability of the gas in the wave guide could result in arcing between regions of opposing charge, ultimately causing catastrophic destruction of the system.

Use of a wave guide requires a different concept of transmission of current and voltage because there is no second conductor to complete the electrical circuit. Instead, electromagnetic lines of force circulate across the major dimension of the wave guide with electric lines of force oriented across the minor dimension, inducing propagation along the length of the line as depicted in figure 3-8. When the open end of a wave guide is properly flared, an electromagnetic horn is created.

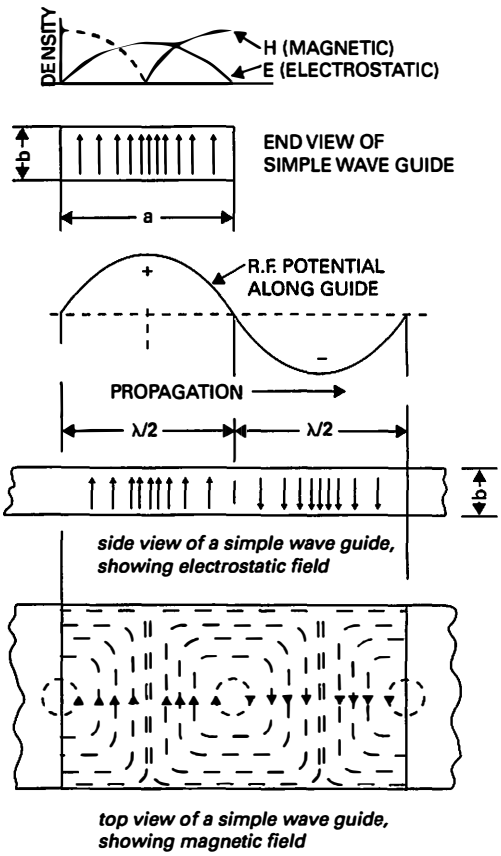


FIGURE 3-8 *Wave guide field diagram*

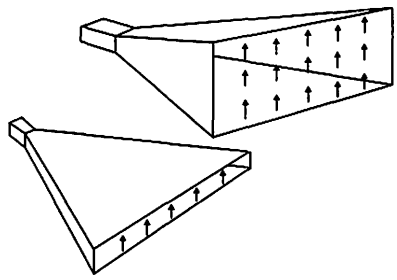


FIGURE 3-9
Electromagnetic feed horns

These feed horns can be employed to release electromagnetic radiation directly into space; see figure 3-9. An RF transparent cover is placed over the open end to maintain gas pressure.

Performance Factors

A detailed investigation is now given of certain factors that affect radar performance. Such factors are pulse shape, pulse width, pulse repetition frequency, carrier frequency, beamwidth, scan rate, antenna design, receiver requirements, bandwidth, and the power relation between peak and average power. The choices of these constants for any particular system will depend upon the tactical use, required accuracy and range to be covered, practical physical size limit, and the problem of generating and receiving the signal.

Pulse Shape and Width

The shape of the pulse determines range accuracy, minimum range, and resolution or definition of the target. Pulse shape is of little importance in detection of a target as long as there is sufficient energy in the return pulse to cause a detectable echo at the receiver. However, the shape of the pulse is very important with regard to range measurements.

The ideal pulse would have perfectly vertical leading and trailing edges. To achieve such a pulse, however, would require an infinite bandwidth (number of frequencies). This is because a square pulse would contain the carrier frequency plus all the odd harmonics of that frequency. In actual practice, rise time—the time for the pulse to reach maximum amplitude—of 0.05 μsec is commonplace. Steepness of the leading edge is a predominant factor in the accuracy of range determination. Slope of the trailing edge causes the receiver to remain blanked (turned off or not aligned to antenna) longer than necessary, thereby reducing minimum range (R_{min}), performance, and target resolution.

Noise may also affect the shape of the pulse, influencing the accuracy of range measurements. Figure 3-10 illustrates a typical pulse shape, superimposed with a pulse corrupted by noise. The time delay between transmitted pulse to receiver pulse will, in effect, be decreased by Δt, thereby making the target appear closer.

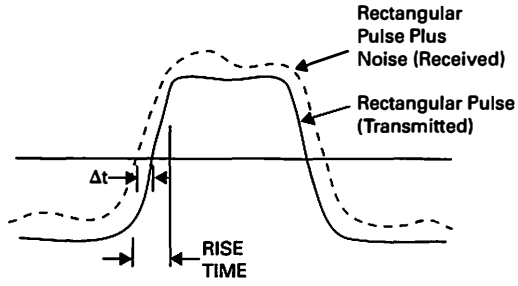


FIGURE 3-10 *Noise-distorted radar pulse*

The width of the pulse determines the radar's range resolution, maximum detection range, and minimum range. Good range resolution, or the ability to separate two targets at nearly the same range, requires a narrow pulse width. To calculate radar range resolution, use the radar range formula:

$$R_{res} = R_{min} = \frac{c(PW)}{2} \qquad (3\text{-}5)$$

where:

R_{min} = minimum range
R_{res} = range resolution (without the pulse compression techniques discussed in chapter 4)
PW = pulse width

Equation 3-5 determines the actual length of the pulse as it propagates through the medium, from leading edge to trailing edge. Figure 3-11 illustrates that when two targets are separated by a distance less than the range resolution, the leading edge of the pulse will be striking the farthest target while the trailing edge of the pulse is approaching the closest target. As the two pulses return to the radar, the leading edge of pulse *B* is hidden in pulse *A*, hence effectively masking the second target.

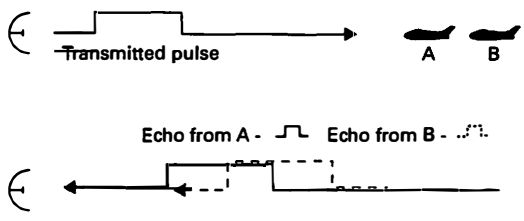


FIGURE 3-11 *Large range resolution large causing overlapping return signals*

Since the radar timer stops when the leading edge of the returning echo is detected and both pulses overlap, only one target would be displayed on the radar scope. For example, a pulse width of 5 μsec means a range resolution of 750 meters. Narrowing the pulse width to 0.1 μsec indicates a range resolution of 15 meters. Thus the narrower the pulse, the better the resulting range resolution.

From the foregoing, it would seem that an extremely narrow pulse width would be desirable. This is not always the case, however. In order to be detected, a target must return an echo that is of sufficient power to be detected by the receiving equipment. The energy in the returned echo may be increased by increasing the peak transmitted power or by increasing the pulse width; however, the peak power that a radar transmitter may produce is limited by the size and quality of its power supply, power amplifier, and other components. Therefore, it is usually more practical to design a radar with a lower peak power and increase the pulse width to maintain a sufficient total energy level. The important criterion in the detection of echoes is the amount of energy in the radiated pulse.

Another factor affected by pulse width is the minimum range. As previously stated, if a target is so close to the transmitter that its echo is received before the transmitter is completely off, the echo reception will not be processed. For good detection of close targets, small pulse widths are necessary, but the limitations mentioned above must also be considered.

Pulse Repetition Frequency

The pulse repetition time (PRT) largely determines the maximum range of the radar set. If the period between successive pulses is too short, an echo from a distant target may return after the transmitter has emitted another pulse. This would make it impossible to tell whether the observed pulse is the echo of the pulse just transmitted or the echo of the preceding pulse. Such a condition is referred to as range ambiguity.

Although the pulse repetition frequency must be kept low enough to attain the required maximum range, it must also be kept high enough to avoid some of the pitfalls a

single pulse might encounter. If a single pulse were sent out by a transmitter, atmospheric conditions might attenuate it; the target might not reflect it properly; or moving parts, such as a propeller, might throw it out of phase or change its shape.

Thus, information derived from a single pulse is highly unreliable. By sending many pulses, one after another, the probability of detection is substantially increased. Radar equipment is therefore designed in such a way that many pulses (ten or more) are received quickly from a single object on a bearing. In this way, the effects of transmission anomalies and pulse-to-pulse variations are averaged out.

Tactical employment of a radar system determines, to a large degree, the PRF to be used. Long-range search radars require a pulse rate slow enough to allow echoes from targets at the maximum range to return to the receiver before the transmitter is again pulsed. Higher pulse rates are used in aircraft interception radars where the maximum range is less because of power limitations.

The following equation will readily determine either the PRF or the maximum unambiguous range (R_{unamb}) if one of them is known:

$$R_{unamb} = \frac{c(\text{PRT})}{2} = \frac{c}{2(\text{PRF})} \qquad (3\text{-}6)$$

Thus, a radar with a PRF of 800 hertz could operate to a maximum range of about 187.5 kilometers without range ambiguity.

In theory, it is desirable to strike a target with as many pulses of energy as possible during a given scan. Thus, the higher the PRF the better. A high PRF combined with a narrow pulse improves angular resolution and range rate accuracy by sampling the position of the target more often. As was shown, maximum range is limited by the PRF. Therefore, a compromise is often reached by selectively increasing the PRF when shorter ranges are desired to obtain the needed accuracy of measurements.

Carrier Frequency

Selection of an appropriate carrier frequency is contingent upon several factors, including the directivity and resolution desired and the existing design limitations on electronic equipment.

The first priority in selecting the carrier frequency is ensuring that the transmitter has the capability to transmit a coherent waveform at that frequency. Coherent waves have two major advantages over noncoherent waves when employed in radar systems. First, Doppler shift due to moving targets can be measured. Second, radar receivers with superior signal-to-noise ratios (SNR) can be used. Radars of older design used noncoherent signals, since at the time of their development, the technology was not available to produce a high-powered coherent signal. Most radars designed and built today are of the more advantageous coherent type.

Directivity is another design factor to consider when choosing carrier frequency. For quasi-optical antenna systems (reflectors), the higher the frequency, the shorter the wavelength and therefore the smaller the antenna required. Conversely, for an antenna of fixed dimensions, the directivity can be increased by using higher frequencies. Higher frequencies also provide for increased target resolution and enable the detection of much smaller-sized targets because of the shorter wavelength.

Atmospheric losses are an additional design criteria for the carrier frequency. The disadvantages of high frequencies include greater propagation loss and the inherent difficulties of generating and amplifying this RF energy. Frequencies from 100 to 20,000 MHz are generally used with the majority of the higher frequencies in aircraft and missile fire control radars. Past and present designations for various radar frequency bands are shown in figure 3-12.

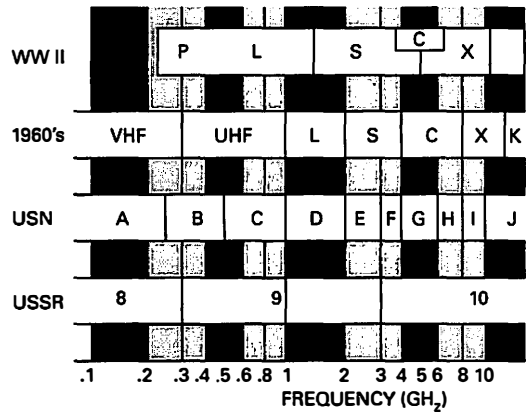


FIGURE 3-12 *Historical radar frequency band designations*

Beamwidth

Tactical application of the radar determines the requirements for the beam. The type of beam needed for target search is different from that required for tracking or for in-close intercept applications. Because the purpose of an early warning radar is to search for and detect targets at maximum ranges, high transmitter power, wide beam patterns, and scan rates that are consistent with the area

under surveillance are the primary requisites. Accuracy of information output is sacrificed in this case for maximum detection probability.

When a radar is used for tracking or guidance, operations that require extreme accuracy, the radar beam must have a narrow width propagated at higher frequencies and rotated in chosen segments of the entire search area. The narrowness of the antenna beam determines the accuracy with which the radar can measure bearing and elevation. It also determines the angular resolution of the radar.

In order to obtain the greatest accuracy in determining bearing and elevation angles, the beam must be as narrow as possible. As shown in figure 3-13, the relative signal strength is a function of angular position of the target from the lobe axis. The maximum signal is received when the axis of the lobe passes through the target.

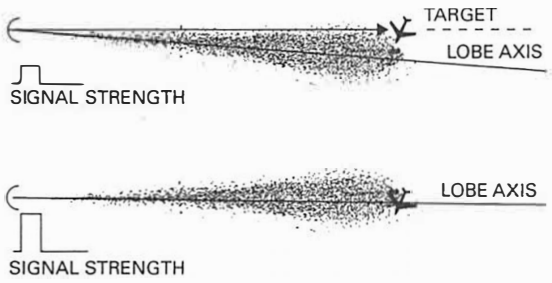


FIGURE 3-13 *Relationship between beam axis and target signal strength*

Measurement of the direction of a target from the radar system is usually given as an angular position. The angle may be measured from true north or with respect to the heading of a ship or aircraft containing the radar set. The angle at which the echo signal returns is measured by using the directional characteristics of the radar antenna system.

The simplest form of antenna for measuring bearing is one that produces a single-lobe pattern. The system is mounted so that it can be rotated. Energy is directed across the region to be searched, and the beam is scanned in bearing until a return signal is detected. The position of the antenna is then adjusted to give a maximum return signal, which then indicates the angular position of the target.

Figure 3-14 shows the receiving pattern for a typical radar antenna. Signal strength is plotted against angular position of the antenna with respect to the target. A maximum signal is received only when the axis of the lobe passes through the target.

As stated above, the angular resolution characteristics of a single-lobe radar are determined by the antenna beamwidth. Classic electromagnetic diffraction theory is

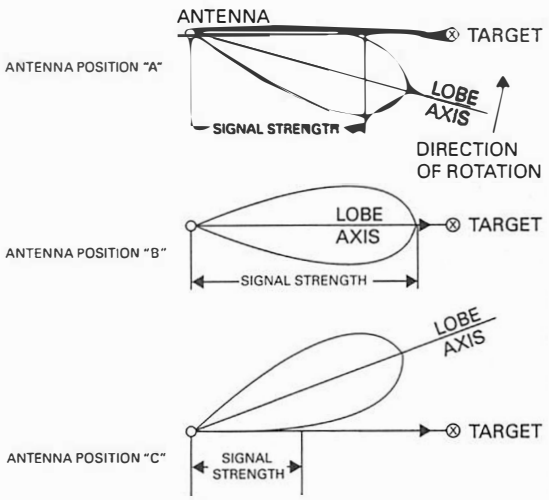


FIGURE 3-14 *Signal strength versus target position in the beam; as the beam sweeps from A to C, signal strength changes with target's position in the beam.*

used to establish the relationship between the physical size of the antenna aperture and its half-power points (3dB down) radiation pattern; see figure 3-15. Therefore, for a diffraction-limited antenna, the beamwidth (ϕ_{-3dB}) would be:

$$\phi_{-3dB} = \frac{k\lambda}{l} \qquad (3\text{-}7)$$

where:

l = antenna dimension in the plane of the measured beamwidth

λ = the wavelength

k = constant of proportionality depending on distribution of the electromagnetic energy across the aperture

A k value of $4/\pi$ is sometimes used for an intermediate or typical pattern. For a linear array or rectangular antenna, a good approximation for k is 0.88, and for a circular aperture the −3dB beamwidth k value is 1.02.

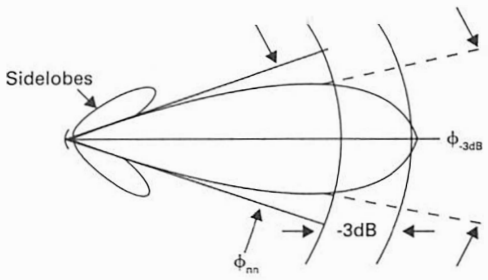


FIGURE 3-15 *Beamwidth is routinely measured between the 1/2 powerpoints (−3dB) in contrast to beamwidth of the null to null (ϕ_{nn}).*

A conflicting characteristic arises with narrow beamwidths for good angle resolution. While operators are usually focused solely on the main antenna beam, there are also "leaks"—more commonly known as lobes or sidelobes—around other axes of the antenna. The formation of sidelobes is a function of antenna design and power output. Typically as the beamwidth becomes narrower, by decreasing k the sidelobes become more pronounced. This can be a serious problem, which will not be discussed here, but it indicates that there are limits on beamwidths for a particular antenna. In general, all antennas will have some sidelobes, but proper antenna design will limit them.

Scan Rate

Another factor that enters into the radar PRF is the rate of angular motion of the antenna in searching extensive regions. If the antenna is moved through too large an angle between pulses, not only will the number of pulses per target be too low, but there may even be volumes of space in which targets may exist without their being detected. In this connection, still another factor to consider is the narrowness of the antenna beam. A narrow beam must be pulsed more often than a wide beam to avoid skipping over targets.

The following equation shows the relationship of the antenna beamwidth, PRF, and antenna scan rate:

$$N = \frac{\phi_{-3dB}\text{PRF}}{\Omega_S} \tag{3-8}$$

where:

ϕ_{-3dB} = antenna beamwidth in radians
Ω_S = antenna scanning rate in radians per second
N = number of pulses returned from a point target as the antenna scans through its beamwidth. (Generally, at least ten pulses must be returned from a point target for reliable detection probability.)

Antenna Gain

A very important characteristic of any radar system is its antenna gain. It is a measure of the ability of an antenna to concentrate energy in a particular direction; see figure 3-16. Two different but related definitions of antenna gain are the directive gain and the power gain. The former is sometimes called the directivity, while the latter is often simply called the gain. The directive gain is descriptive of the antenna radiation pattern, whereas the power gain is an indication of antenna efficiency.

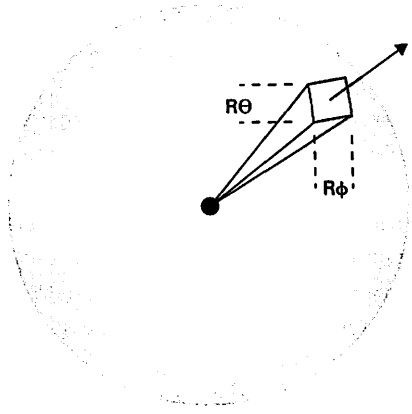


FIGURE 3-16
Spherical spreading from a point source

Directive gain of a transmitting antenna may be defined as:

$$G_D = \frac{\textit{maximum radiation intensity}}{\textit{average radiation intensity}} = \frac{\textit{maximum power per steradian}}{\textit{total power radiated isotropic}} \tag{3-9}$$

As one might expect, the directive gain and the beamwidth have an inverse relationship to each other. If θ_{-3dB} and ϕ_{-3dB} are the bearing and elevation half-power beam widths, respectively, then the relationship is

$$G_D = \frac{\frac{P}{R\theta_{-3dB}R\phi_{-3dB}}}{\frac{P}{4\pi R^2}} = \frac{4\pi}{\theta_{-3dB}\phi_{-3dB}} \tag{3-10}$$

where:

θ_{-3dB} and ϕ_{-3dB} are measured in radians
P = power radiated (watts)
R = radius from source (meters)

EXAMPLE 3-2: For a 3° beamwidth (0.052 radian) in both bearing and elevation, what is the directive gain?

SOLUTION:

$$G_D = \frac{4\pi}{(0.052)^2} \cong 4{,}600$$

The definition of directive gain is based primarily on the shape of the radiation pattern. Power gain, which will be denoted by G, includes the effects of antenna losses and any other loss that lowers the antenna efficiency.

Power gain is defined as:

$$G = \frac{\text{maximum radiation intensity from the antenna of interest}}{\text{radiation intensity from omnidirectional source with same power input}} \tag{3-11}$$

Because of losses, power gain is always smaller than the directive gain.

It should be noted that although power gain of an antenna may exceed unity by a considerable amount (because of the way power gain is defined), the efficiency of the antenna is always less than 100%.

The size of the antenna aperture also determines its effectiveness as a collector of energy. The amount of power available to the receiver is a function of the energy density of the returning echo (watts per m^2) and the effective area of the antenna in m^2. The antenna's effective area or aperture (A_e) is related to the carrier frequency of the radar, the construction of the antenna, and the antenna's physical size or area, A. The difference between the physical area of the antenna and the effective aperture is a measure of its efficiency, typically 65%–85%. Thus:

$$A_e = \rho_a A \tag{3-12}$$

where:

ρ_a = antenna efficiency
A = physical area of the antenna (m^2)
A_e = antenna aperture in m^2

Antenna aperture and power gain (G) are then related, and substituting equation 3-7, noting that $l^2 = A$ for a square antenna such that vertical and horizontal beamwidths are the same, to equation 3-10 and applying equation 3-12 results in:

$$G = \frac{4\pi\rho_a A}{k^2\lambda^2}$$

or

$$G = \frac{4\pi A_e}{k^2\lambda^2} \tag{3-13}$$

Notice that antenna gain, and thus directivity, increases as wavelength decreases for a fixed antenna size. Stated another way: for a specific frequency, increased directivity (narrower beam width) requires a larger antenna.

Receivers

Although the radar antenna transmits powerful pulses of directional electromagnetic energy, only a small part of this energy actually strikes a distant target. Furthermore, since the return energy from the target is reflected over a wide angle, only a minuscule fraction of the transmitted energy returns to the radar antenna. Thus, the function of the radar receiver is to receive weak target returns from the antenna and amplify them sufficiently to provide desired target information. Further, the weaker the signal that the receiver is capable of using, the greater the effective range of the radar set.

It is the steady improvement in receiver design that has made a greater contribution to the usefulness of radar than any other single development. Present-day receivers are extremely sensitive, having been perfected to the point where they will discriminate signals on the order of a picowatt (10^{-12}) and will amplify and display them in many useful ways.

The absolute lower limit to the sensitivity of a radar receiver is set by a phenomenon called noise. Noise consists of any of several unwanted voltage inputs to the receiver that may ultimately obscure a weak target return signal. Noise may be attributed to many external sources, ranging from atmospheric disturbances to deliberate attempts made by opposing forces to saturate or jam the return radar signal. These are transitory in nature, however, and are less critical to the determination of system performance than the noise generated by the circuitry of the system itself.

Noise may be attributed to internal noise because in all electronic systems, there will always be some noise present. If nothing else, the fact that systems are not operated at absolute zero (–273°C) guarantees that there will be random electron motion in all electronic components. Electron movement constitutes a current, and this random current through any resistance produces a noise voltage in the system. Both external and internal noise combines to create the overall noise levels for the system.

The background noise is specified by its average value, called the noise-equivalent-power (NEP). This directly equates the noise to a detected power level so that it may be compared to the return. Successful detection of a target requires that

$$P_r > \left(\frac{S}{N}\right)\text{NEP} \tag{3-14}$$

where P_r is the power of the return signal. Since the combination of terms on the right side of the equation determines the radar system's performance, it is given a unique designation, S_{min}, called the minimum signal for detection.

$$S_{min} = \left(\frac{S}{N}\right)\text{NEP} \tag{3-15}$$

Since S_{min}, expressed in watts, is usually a small number, it has proved useful for defining the decibel equivalent, MDS, which stands for minimum discernible signal:

$$\text{MDS} = 10\log\left(\frac{S_{min}}{1\,\text{mW}}\right) \tag{3-16}$$

When using decibels, the quantity inside the brackets of the logarithm must be a number without units. In MDS,

this is done by using the fraction S_{min}/1 mW. The special notation dBm is used for the units of MDS, where the "m" stands for 1 milliwatt. This is shorthand for decibels referenced to 1 mW, which is sometimes written as dB/1 mW.[ii]

A radar receiver has a means of setting the "threshold level," a level that a signal must exceed to be displayed on an indicator or to be used in some other manner. Unfortunately, noise signals are not at one constant level and can vary, as illustrated in figure 3-17. If a threshold level is set high enough to prevent any noise signals from appearing, the radar's sensitivity is reduced considerably. If the threshold is set too low, then many false indications are given, and these noise signals tend to complicate detection.

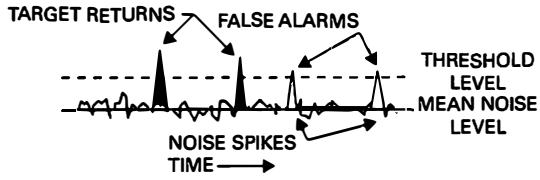


FIGURE 3-17 *Comparative signal amplitudes and threshold levels*

Therefore, a compromise threshold level is selected by adjusting manually or automatically for specific operating conditions. The threshold level then sets the "false-alarm rate" of the radar, which is simply the number of false-alarm signals that appear in a given time period. In summary, the probability of a false-alarm signal occurring depends upon the threshold setting.

The ratio of the return radar signal input to the noise voltage is called the signal-to-noise ratio (SNR). Some receivers continuously monitor the background noise and automatically adjust the threshold level to maintain a constant false-alarm rate (CFAR), and therefore are called CFAR receivers. Some other common receiver features used to reduce the randomness or pull the signal out from the noise are:

1. Sensitivity time control (STC), which is a signal-processing technique that minimizes the impact of returns caused by sea state. After a pulse is transmitted, the system's SNR is increased, thereby causing the spurious signals from the sea not to be displayed; after a short period the SNR is gradually reduced until the noise limit is reached. Using STC enables reduced clutter around the transmitter but provides maximum sensitivity to the receiver for detection of long-range weak returns.
2. Pulse integration is a technique that uses the random nature of noise and the steady target signal to achieve detection. Assume that a target is at a very long range and has a weak signal. This signal is so weak that it is below the SNR threshold. Adjusting the receiver gain will not solve this problem because the noise will be amplified as much as the signal and the SNR will remain the same. Pulse integration solves this problem by summing successive pulses along the same bearing. Since noise is random, noise power averages to a mean level and the signal level from the target increases to a point of detection.
3. Fast time constant (FTC) is a signal-processing technique that reduces the effects of long-duration returns caused by weather. Weather events such as rain slowly increase in amplitude. The processor uses a differentiator to determine the rate of signal change and filters out returns that rise slowly. Returns from a target, however, appear as a spike rising and falling rapidly.

Bandwidth

As mentioned previously, the pulse-modulated carrier waveform consists of a series of sine waves having odd harmonic relationships that combine to form the nearly square pulse amplified by the transmitter. Thus, though the preponderance of energy transmitted by the radar is at the intended frequency, the process of achieving nearly vertical leading and trailing edges results in inclusion in the pulse of waveforms other than the carrier frequency. The receiver must therefore be designed to detect and amplify a band or range of frequencies above and below the transmitted frequency.

This bandwidth (BW) must be sufficiently large to accommodate all frequencies included in the transmitted pulse, or the pulse shape will be distorted by the receiver and degradation of performance will result. The presence of Doppler shift in returning echoes further increases the bandwidth requirements, whether or not the radar includes Doppler-processing capability.

Typical bandwidths for radar receivers range from 1 to 5 megahertz. The gain of the receiver should be as uniform as possible over the entire bandwidth. However, if the receiver bandwidth is too wide, too much noise will be amplified. The result is that the minimum discernable signal is increased. A high signal-to-noise ratio is therefore required to ensure that the receiver can detect a weak signal from a distant target and override the noise amplified in the receiver.

Thus, the amount of noise that will pass through the receiver is a function of the bandwidth of the receiver. In general, a reduction of the bandwidth of the receiver reduces the noise in the output, but does so at the expense of distortion of the pulse shape. The amount of distortion that can be tolerated is the limiting factor in increasing the signal-to-noise ratio by reducing the bandwidth.

Through Fourier analysis it can be shown that the null-to-null bandwidth and pulse width are inversely proportional, given by:

$$BW_{nn} = \frac{2}{PW} \quad (3\text{-}17)$$

where:

BW_{nn} = null-to-null range of frequencies
PW = time that the transmitter is on

This relationship intuitively states that as PW increases, the BW transmitted and consequently returned in the pulse decreases. Figure 3-18 compares two different pulse widths, both in the time and frequency domains to illustrate the point.

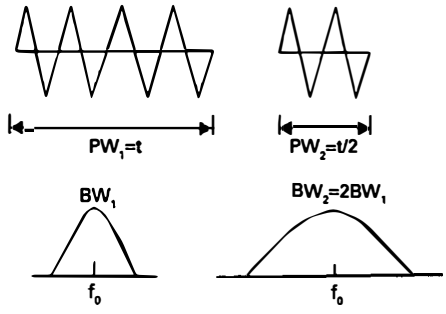


FIGURE 3-18 *The narrower the pulses, the wider the central frequency bandwidth*

The –3dB bandwidth point is approximately one-half the null-to-null value or

$$BW_{-3dB} = \frac{1}{PW} \quad (3\text{-}18)$$

Transmitter Power

Greater transmitter power improves detection range, up to a point. The amount of energy a radar transmits is equal to the power of the transmitter times the length of time the radar is transmitting. The more energy transmitted, the greater the return signal strength and subsequently the higher the probability of detection.

However, there is a maximum limit to the intensity of the transmitted signal. The peak power determines the intensities of the electromagnetic fields along the transmission path. If these fields are too powerful, unacceptable losses of power and the potential for equipment damage will occur.

The two effects from large EM fields are corona and arcing. Corona occurs when the electric field is so strong that it ionizes air, and arcing occurs when this ionized air develops into a conductive path. These two effects provide a maximum limit to the amount of peak transmitter power.[iii]

The heating of transmitter components as well as space and weight concerns form other limits to transmitter power. The greater the average transmitter power, the more robust the equipment, and consequently, the more heat generated by the components. This heat must be transferred from the equipment, thus requiring more complex and extensive cooling systems as power levels increase.

Radar Cross Section of Targets

The radar cross section (RCS) of a target, denoted by the symbol σ, is defined as the area intercepting that amount of power which, when scattered equally in all directions, produces an echo at the radar.

The radar cross section is, of course, a characteristic of the target and as such is not actually one of the radar performance factors. However, it is one of the major factors affecting the strength of a radar return. For most common types of radar targets such as aircraft, ships, and terrain, the radar cross section does not bear a simple relationship to the physical area, except that the larger the target size, the larger the cross section is likely to be.

When an object is struck by an electromagnetic wave, a portion of the incident energy is absorbed as heat, and the remainder is reradiated (scattered) in many different directions. Major factors determining the radar cross section are target size, shape, skin material, aspect angle, polarization, and radar carrier frequency.

As an example of how much just one of these factors can change the radar cross section, consider a change in aspect angle where the percentage on incident energy directed back can vary by as much as a factor of 1,000. There is no simple formula that will give the cross section of an actual target such as a ship or aircraft. Thus, cross-sectional data are experimentally obtained, and the results are valid only for a specified frequency and aspect of the target.

Remember that as the wavelength approaches the size of the target and then becomes smaller, the waves begin to get more reflection off the target. That is why some radars can see through rain, and weather radars adjust the wavelength to actually find the rain. Here is a

simplification of the relationship of wavelength, or frequency, to target size:

- *Rayleigh region:* If the target is a lot smaller than the wavelength of the radar system, the target is said to be in the Rayleigh region. If the target is in the Rayleigh region, the radar cross section of the target tends to be smaller than the target's physical size.
- *Resonance region:* If the target is of similar dimension to that of the wavelength, the target is said to be in the resonance region. In the resonance region, the radar cross section of the target may vary a great deal but tends to be larger than the physical size of the target.
- *Optical region:* The optical region occurs when the target is much larger than the operating wavelength of the radar. This is quite often the case with operational radar systems whose wavelengths are normally on the order of centimeters in length. When operating in this region, the radar cross section of the target is similar to its physical size. In the optical region where the wavelength is small compared to the object's dimension, the shape of a scattering object has far greater effect on its RCS than does its physical size.

The implication of these three regions is that the operating wavelength should not be selected in total isolation from target considerations.

Figure 3-19 is a polar representation of the radar cross section of a typical aircraft. Notice the very large changes in cross section that occur as the view angle changes. Nulls and maxima are closely spaced. The variations shown are of greater magnitude than one may realize at first.

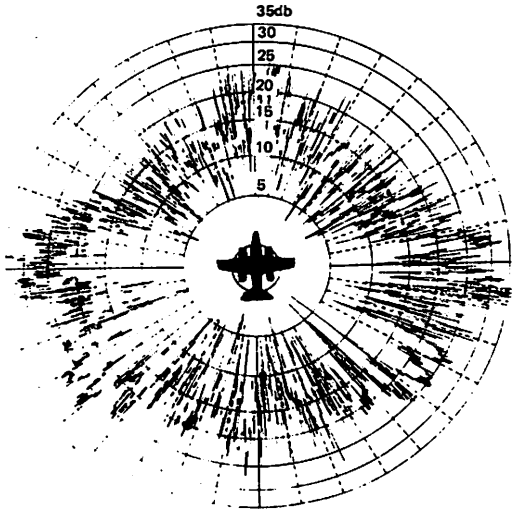


FIGURE 3-19 *Experimental cross section of a tactical bomber-sized aircraft as a function of azimuth angle*

Inspection of the scale in figure 3-19 reveals an overall change of approximately 33dB. Recall that every 3dB is twice the power, so a 33dB change is the equivalent of a power ratio of 2,000! Observe, also, that the relative size of the radar cross section obtained from a broadside view of the aircraft is much larger than that corresponding to other views. Two factors are responsible for this. When viewed at 90°:

1. The aircraft appears physically larger.
2. The junction of wings and fuselage functions as multiple corner reflectors; see figure 3-20.

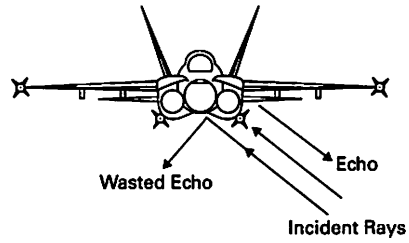


FIGURE 3-20 *Illustration of the manner in which the junctions of wings and fuselage, and curved surface of aircraft, modify radar cross section*

Table 3-1 gives rough approximations of radar cross sections for various types of targets at the optimum radar frequency.

TABLE 3-1 *Examples of radar cross sections in radar frequency bands*

Target	Square Meters
Conventional winged missile	0.1
Small, single-engine aircraft	1
Small fighter or four-passenger jet	2
Large fighter	6
Medium bomber or medium jet airliner	20
Large bomber or large jet airliner	40
Jumbo jet	100
Helicopter	3
Small open boat	0.02
Small pleasure boat (20–30 ft)	2
Cabin cruiser	10
Ship	Displacement tonnage in m^2

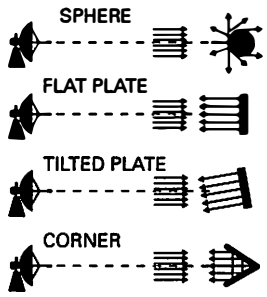


FIGURE 3-21 *Backscatter from different shapes*

Table 3-2 can be used to review the major characteristics of radar performance factors.

TABLE 3-2 *Summary of radar performance factors*

Factor	Radar Characteristics Affected
Pulse shape	Range accuracy; range resolution; minimum range
Pulse width	Range resolution; minimum range; maximum detection range
Pulse repetition frequency	Maximum unambiguous range; range rate accuracy; angular resolution; detection probability
Carrier frequency	Directivity; target resolution; propagation loss; size of equipment
Scan rate and beamwidth	Probability of detection; angular resolution
Receiver sensitivity	Maximum detection range
Transmitter power	Maximum detection range; physical size
Antenna gain	Maximum detection range
Radar cross section	Maximum detection range

The Simplified Radar Range Equation

Several of the aforementioned factors affecting radar performance can now be combined to develop the basic radar range equation.

If peak power output of a radar transmitter is denoted by P_{peak} and is radiated uniformly in all directions, the power density (power per unit area) at any distance R from the radar can be determined by dividing the transmitted power by the surface area of an imaginary sphere of radius R.

$$\text{power density for an omnidirectional antenna} = \frac{P_{peak}}{4\pi R^2} \tag{3-19}$$

It has been shown that radars employ directional antennas rather than omnidirectional ones. The power output of a directional antenna is related to P_{peak} by the power gain.

$$\text{power density for a directional antenna} = \frac{P_{peak} G}{4\pi R^2} \tag{3-20}$$

The above relationship defines the power that will reach a target at a distance R from the radar. The power that is reradiated by the target is a function of the radar cross section of the target, σ.

$$\text{power reradiated by target} = \frac{P_{peak} G\sigma}{4\pi R^2} \tag{3-21}$$

The signal undergoes spherical spreading again on its way back to the radar receiver. The power density of the return pulse at the radar antenna is now:

$$\text{power density of return at radar} = \frac{P_{peak} G\sigma}{(4\pi R^2)^2} \tag{3-22}$$

Only a small amount of return pulse from the contact is collected by the radar antenna and sent to the receiver. The effective antenna aperture (A_e) is used to determine the actual amount of returned pulse received. Therefore, the power density at the radar receiver P_r is now:

$$P_r = \frac{P_{peak} G\sigma A_e}{(4\pi R^2)^2} \tag{3-23}$$

where:

P_r = power returned
P_{peak} = power transmitted by radar
G = antenna power gain
A_e = antenna effective area
σ = radar cross section of target
R = distance from radar site to target

When the minimum signal for detection (S_{min}) is equal to the power density at the radar receiver, target detection will result. Therefore,

$$S_{min} = P_r = \frac{P_{peak} G\sigma A_e}{(4\pi R^2)^2} \tag{3-24}$$

Solve for range and refer to this range as the maximum range (R_{max}):

$$R_{max} = \left[\frac{P_{peak} G\sigma A_e}{(4\pi)^2 S_{min}}\right]^{1/4} \tag{3-25}$$

Therefore, the "simplified" radar range equation is:

$$R_{max} = \sqrt[4]{\frac{P_{peak} G A_e \sigma}{(4\pi)^2 S_{min}}} \tag{3-26}$$

Equation 3-26 is the simplified radar equation derived by considering only peak transmitter power. A more detailed analysis of the factors determining radar range would have shown that the maximum range of a pulse radar is dependent on the amount of energy contained in a pulse. Intuitively, it makes sense to say that if a radar transmits a greater amount of energy, the maximum range at which a target may be detected will be greater.

The one-fourth power relationship occurs because the signal is twice subjected to the inverse square law relationship that governs free-space transmission of electromagnetic waves having spherical spreading. The relationship applies first to power transmitted in the forward path and again in the return path. This fourth-root relationship between power and maximum range is the explanation of why long-range radars require such powerful transmitters. The following numerical example will illustrate this.

EXAMPLE 3-3: A 100-kilowatt radar can detect a specified target at a maximum range of 100 kilometers. How much transmitter power would be required to detect the same target at 500 kilometers (neglecting the effect of earth curvature)?

SOLUTION:

$$\frac{R_2}{R_1} = \sqrt[4]{\frac{P_2}{P_1}}$$

$$P_2 = \left(\frac{R_2}{R_1}\right)^4 \qquad P_1 = \left(\frac{500\text{ km}}{100\text{ km}}\right) \times 100\text{ kw} = 62.5\text{ Mw}$$

The 62.5-megawatt transmitter does not exist. Making this possible would necessitate significant improvements in the efficiencies of transmitter and wave guide components to avoid serious losses in this part of the system. Evidently, extending the range of this radar to 500 kilometers, even with the aid of a more practical 1/2- to 2-megawatt transmitter, would require vast improvements in receiver gain, or large increases in antenna gain and/or effective area. Improved receiver gain implies maximum effort in noise reduction. It should now be clear why it is so important to reduce the noise to as low a level as technologically possible. Also implied is that the cost of a radar is exponentially proportional to the range required.

Use of the Simplified Radar Range Equation

This expression of the radar range equation does not include terms for the values of the various losses associated with specific equipment, nor does it completely account for applicable statistical theory associated with detection. The term S_{min} assigns a detection probability of 0.5 to a situation where the returning echo power at the receiver (first preamplifier) just equals the noise level. Although this chapter was not intended to provide sufficient range prediction accuracy for engineering purposes, it can provide to the tactician significant data concerning sensor performance where the actual detection range has been experimentally established for a target of known radar cross section (σ).

Under conditions where the actual detection range has been established and a new range is to be determined for a given change in target size, gain, antenna size, power, or S_{min}, the terms for the nonincluded variables can be assumed to drop out when using ratio and proportion methods of problem solving. The assumed accuracy of the range determined in this way is sufficient for rough tactical decision making when used for radar in the context of the equipment manufacturer's instructions and directives from higher authority.

EXAMPLE 3-4: Recent fleet exercises have shown that your Airborne Early Warning (AEW) aircraft radar can detect a target drone ($\sigma = 1\ m^2$) at a range of 75 kilometers. At what range could a cruise missile be detected by the same radar if its radar cross section is $0.5\ m^2$?

SOLUTION:

$$R_{unamb} = \frac{c(\text{I}}{}$$

$$R_2 = \sqrt[4]{\frac{\sigma_2}{\sigma_1}} \times R_1 = \sqrt[4]{\frac{0.5\text{ m}^2}{1\text{ m}^2}} \times 75\text{ km} = 63\text{ km}$$

EXAMPLE 3-5: During operations in the Mediterranean Sea, the on-station E-2C detects a large contact to the north at 400 kilometers with the AN/APS 125 radar (peak power 1 Mw). A power fluctuation causes the low-voltage protection circuits to shut down the radar. When operation is restored, it is discovered that the MDS is degraded by 3dBm, and peak power is reduced to 750 kilowatts. At what range will the target now be detected?

SOLUTION:

$$R_2 = \sqrt[4]{\frac{\frac{P_{peak\,2}}{S_{min\,2}}}{\frac{P_{peak\,1}}{S_{min\,1}}}} \times R_1$$

MDS degraded by 3dBm means $S_{min\,2} = 2 \times S_{min\,1}$.

$$P_{peak\,1} = 0.75\ P_{peak\,2}$$

$$R_2 = \sqrt[4]{\frac{\frac{0.75}{2}}{\frac{1}{1}}} \times 400\text{ km} = 313.02\text{ km}$$

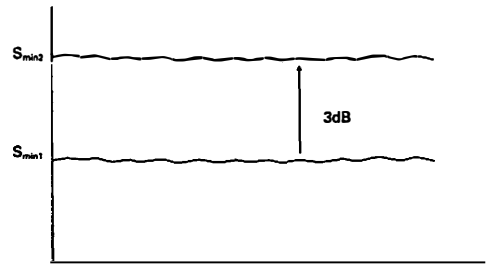


FIGURE 3-22 *3db degradation and radar performance*

Environmental Considerations

When operating high-energy RF sources, care must be exercised because of the possibility that mutual influence among various transmitters and receivers will cause signal degradation. In many cases, signals from other similar systems can constitute a serious interference problem. This interference can be divided into two broad categories: antenna-to-antenna coupling and nonantenna coupling.

Antenna-to-Antenna Coupling

High levels of energy coupled into a receiving antenna pose a severe interference. In addition, reception of these high-energy signals can cause degradation or burnout of crystal detectors.

Non-Antenna Coupling

Anomalous detection Anomalous detection of RF signals can occur in a variety of equipment: public address systems, motion picture sound systems, etc. In addition, lifelines and other metallic standing rigging aboard ships can act as receiving antennas for electromagnetic energy. If not properly grounded, or if corrosion results in poor metal-to-metal contact, arcing can occur, which degrades radio communications and could result in an electrical shock hazard.

Hazards to ordnance (HERO) RF energy coupled into the circuit wiring of electrically fired ordnance can create fields powerful enough to cause premature ignition of rocket motors or dud the weapon. In most cases, specific distance requirements between transmitters and ordnance are specified.

Hazards to personnel (RADHAZ) Human tissue absorbs the energy present in these RF waves with a resultant temperature rise. Frequency ranges between 1 and 3 GHz can result in as high as almost 100% absorption (think of a microwave oven). Below 1 GHz, absorption is less than 40%; however, a new danger exists. Because of the longer wavelength, penetration into the body is deeper. As a result, heating of internal organs occurs, which is not as apparent as the rise in skin temperature. When operating high-energy devices, care must be taken to prevent accidental irradiation of personnel close to the transmitter. Above 3 GHz, more of the energy is reflected, but absorption still accounts for about 40% of this energy.

Summary

Typical pulse-echo radars are made up of seven basic components: transmitter, receiver, power supply, synchronizer, duplexer, antenna system, and display. Although all components share an equal importance to overall radar effectiveness, the antenna system is unique in that it serves in multiple capacities.

First of all, it is both the transmitter and receiver of broadcast energy. In a pulsed system, this dual role is accomplished by the same antenna through the employment of a duplexer to isolate the receiver during the transmit cycle.

Secondly, the antenna design directly determines the shape of the radar beam and thus the accuracy of target positioning. This shaping is accomplished through design applications that incorporate electromagnetic waves for reflection, refraction, and interference.

Pulse-range determination is based on the fact that a transmitted radar pulse travels at the speed of light, and therefore range is a function of pulse round-trip time. The accuracy of target bearing information is based on sensitivity to changes in return signal strength. The more accurately these changes can be calibrated, the more accurate the bearing information will be.

Practical limitations on minimum radar return signal detection are set by noise. Noise may be environmental radiation, system noise, or the result of thermally induced electronic fluctuation within the input circuitry of the receiver.

Radar performance is determined by many factors, several of which are closely interrelated. Among these are pulse shape, width and repetition frequency, antenna gain, and target reflectivity. The absolute maximum range of a radar is determined by transmitter power, antenna gain and reception area, target radar cross section, and receiver sensitivity. A change in any of these factors will alter maximum range in accordance with a one-fourth power relationship.

---

[i] Gurvine and Strauss, 6–17

[ii] Hall, 92

[iii] Stimson, 113

4 Radar Systems

Introduction

In the previous chapter, basic principles of radar operation were discussed. The straightforward pulse radar was introduced along with the major factors that affect the signal strength of the return echo. In this chapter, other types of radars are introduced, namely continuous-wave radar, pulse-Doppler radar, MTI, and phased array-type radar systems.

Continuous-Wave (CW) and Frequency Modulated Continuous Wave (FMCW) Radar

Continuous-Wave Radar

The first documented case of a radio wave detecting an object was by Christian Hulsmeyer in May 1904. This demonstration showed that when riverboats were in a beam of continuous radio waves, the radio waves returned from the objects and could be detected.

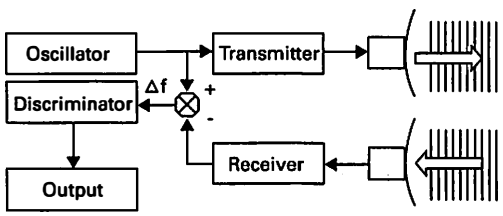


FIGURE 4-1 *CW radar block diagram*

An elementary continuous-wave (CW) radar with a minimum of essential components is illustrated in figure 4-1. This type of radar radiates uninterrupted energy at a fixed frequency. There is no timing circuit of the sort encountered in a pulse radar because there is no pulsing of the signal. Because of the continuous operation, peak power of the radar is equal to its average power; therefore its duty cycle is unity or:

$$\frac{P_{avg}}{P_{peak}} = 1 \qquad (4\text{-}1)$$

Energy from a target is then reflected back to the receiving antenna by objects within the beam.

This type of system uses a separate antenna for receiving since the receiver and transmitter must operate simultaneously. Because there is some hazard that the transmitted energy will directly enter the receiving antenna, the antennas are located some distance apart from each other in order to reduce this problem. If displacement is not possible, energy isolation can be provided by making the antennas highly directional or by using metallic baffles with electromagnetic absorbent material.

The oscillator provides a reference signal that drives the transmitter. The transmitter then amplifies the signal for transmission.

The same reference signal is also provided to the mixer. The return signal for a moving target differs from the sample signal because of Doppler shift. In the mixer, the reference signal and the return signal are compared, determining the difference between the two signals caused by the Doppler shift. The output of the mixer is then a function of this difference, which is amplified and sent to an indicator.

A CW radar is continuously transmitting and receiving; consequently, there is not a minimum detection range, R_{min}. However, because simple CW radars have no time reference such as that found in pulse radars, they are unable to measure range in a conventional manner.

The change in received frequency due to Doppler shift is proportional to the target's velocity in the line-of-sight (LOS). CW radars measure only radial velocity or that component directly inbound or directly outbound.

Radial velocity or velocity in the line of sight, v_{LOS} can be calculated with knowledge of the transmitter/receiver and target velocity, designated as v_1 and v_2, respectively, and the angles between their direction of motion and the

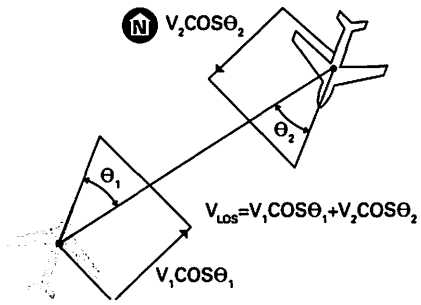


FIGURE 4-2 *Calculating the relative speed in the line-of-sight*

line-of-sight, designated θ_1 and θ_2; see figure 4-2. The combined velocity in the line-of-sight is[i]:

$$v_{LOS} = v_1 \cos\theta_1 + v_2 \cos\theta_2 \qquad (4\text{-}2)$$

where:

v_{LOS} = velocity in the line-of-sight
v_1 = velocity platform 1
v_2 = velocity platform 2
θ_1 = relative bearing from platform 1 to platform 2
θ_2 = relative bearing from platform 2 to platform 1, also known as the target angle

LOS velocity can also be interpreted as the instantaneous rate of change in the range, or range rate. As long as the problem is confined to two dimensions, the angles also have simple interpretations. Since the cosine function is even, it makes no difference whether the angle is positive or negative (strictly speaking in the Navy, relative bearings are always positive, ranging from 0° to 359°). The angle θ_1 is the relative bearing to the target. This is the difference between the course of the transmitter/receiver and the true bearing to the target. To convert between relative and true bearings, use:

relative bearing = true bearing – heading
θ_1 = true bearing of platform 2 – true heading of platform 1

The angle θ_2 is the target angle (relative bearing of transmitter/receiver from target). To determine θ_2, the true bearing of platform 1 from platform 2 must be found. This is simply the reciprocal of the true bearing of platform 2 from platform 1.

The reciprocal bearing is found by:

reciprocal bearing = true bearing ± 180°
θ_2 = true bearing of platform 1 – true heading of platform 2

Again, mathematically it does not matter if this result is positive, negative, or even beyond 360°, although within the realm of the real world the range is limited to 0°–359°.

Assuming that the range rate is known, the shift in returned frequency f_D is[ii]:

$$f_D = \frac{2v_{LOS}}{\lambda} \qquad (4\text{-}3)$$

where:

f_D = Doppler shift (Hz)
λ = wavelength of original signal (meters)
v_{LOS} = velocity in the line-of-sight

In actual practice equation 4-3 is rearranged because the mixer measures the Doppler shift directly, and from that, the velocity in the line-of-sight with respect to the target is determined:

$$v_{LOS} = \frac{\lambda f_D}{2} \qquad (4\text{-}4)$$

Based on operating frequencies and aircraft speeds, Doppler frequencies will normally be less than 50 KHz. For a CW radar to capture positive and negative frequency shifts in this range, it would require a bandwidth of 100 KHz (+/–50 KHz). This is in contrast to pulsed radars, which typically have bandwidths of several megahertz. Because of this decreased bandwidth requirement, CW radars are typically less noise limited, causing a higher signal-to-noise ratio than pulse radars.

However, due to their dual antenna configuration, they are usually power limited. Additional noise reduction can be accomplished through filtering the noise from sea returns, land masses, and other such undesirable targets. Moreover, the amplifier can be tuned in such a way as to select desired frequency bands in the Doppler shift bandwidth so that only the signals from targets of a certain range rate will be amplified.

For example, the amplifier could be tuned to accept only frequencies that could be expected from Doppler shift of inbound targets with a certain rate of closure. The radar would then single out only targets with those characteristics. This gives the radar a target selection and identification capability that is a distinct advantage in some circumstances.

Elementary CW radars as discussed have several inherent disadvantages. One is that without a Doppler shift, $f_D = 0$, the target is not detected. A target with zero Doppler shift is not moving with respect to the transmitter. In a sense, the radar is "blind" to targets with no radial

relative motion. The target either is at the same course and speed as the transmitter or is circling the transmitter. One situation where this "blindness" could occur is in a "tail chase" between two aircraft.

Another disadvantage for CW radars is that they are unable to measure range. The ability to separate moving targets from stationary targets is considered so desirable, however, that ways have been found to modify the more useful form of the pulse radar so that the Doppler effect can be used as well to provide range determination; this is discussed in the next section of the chapter.

Modulation

Radar and communications transmissions are modulated to convey information, to enhance signal processing in the receiver, or to contend with countermeasures of various types. Variation of the amplitude of the carrier wave (amplitude modulation, AM), the frequency (frequency modulation, FM), or the phase (phase modulation, PM) may be employed in CW systems or be superimposed on the output of a pulsed system individually or in combination.

Pulsing the transmitter output is also a means of modulation that can be varied in several ways to provide coding for information security or countermeasures resistance. The duration of a pulse can be varied, as can its position in a pulse train, carrier frequency, or amplitude.

Many types of radar combine several of the above methods as a means of analog or digital communication with a weapon to direct its flight path, to assist its guidance system in target acquisition, or to activate fuzing systems see figure 4-3.

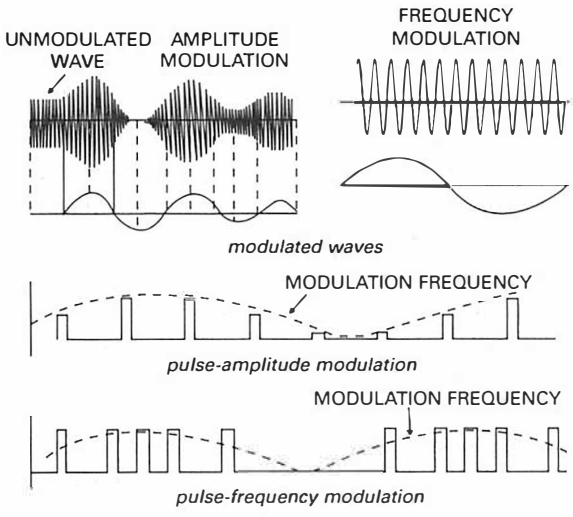


FIGURE 4-3 *Electromagnetic energy modulation techniques*

Frequency-Modulated CW Radars (FMCW)

With a few modifications to basic CW radar, range can be measured directly, much akin to the classic pulse radar. A technique commonly used is frequency modulation. An example of a frequency-modulated output signal is plotted against time in the frequency modulation chart of figure 4-4. As shown, the frequency of the output signal is caused to increase linearly from 420 MHz to 460 MHz and then quickly return to 420 MHz. The sequence begins again at 420 MHz and repeats.

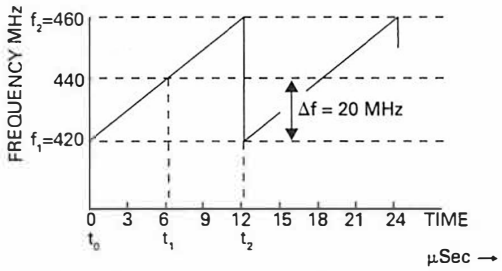


FIGURE 4-4 *Frequency modulation chart*

Since the frequency changes regularly through a range of 40 MHz as a programmed function of time, its value at any instant in the cycle is used as the basis for computing the time elapsed after the start of that sequence. For example, at time t_0, the transmitter sends out a 420 MHz signal that travels toward an object. It strikes the object and returns to the receiver at time t_1. At this time, the transmitter is sending out a new frequency of 440 MHz.

Thus, since frequency changes from 420 to 460 MHz in an interval of 12.0 microseconds, time required for the 420 MHz signal to change from 420 to 440 MHz is 20/40 of 12.0 or 6.0 microseconds, written as:

$$\Delta t = \frac{T\Delta f}{f_2 - f_1}$$

$$\frac{12\,\mu\text{sec}(20\,\text{MHz})}{460\,\text{MHz} - 420\,\text{MHz}} = 6.0\,\mu\text{sec} \qquad (4\text{-}5)$$

In the frequency modulation system, when an echo is received, two separate signals are supplied to the mixer simultaneously. For example, at t_1, both the 440 MHz transmitted signal and the 420 MHz reflected signal are fed to the mixer. The two signals are combined by the discriminator and a beat frequency results. The frequency of the beat note varies directly with the distance

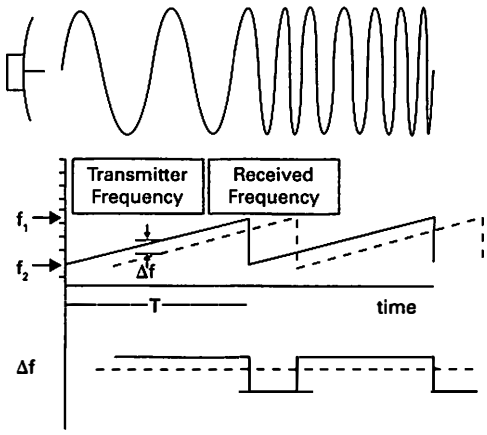


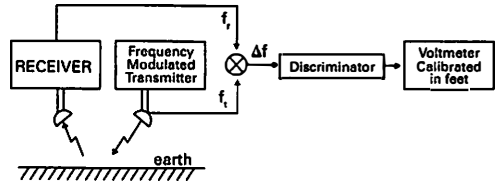
FIGURE 4-5 *FMCW theory, $\Delta f = f_2 - f_1$, resulting in a beat frequency represented by the dotted line*

to the object, increasing as the distance increases; see figure 4-5.

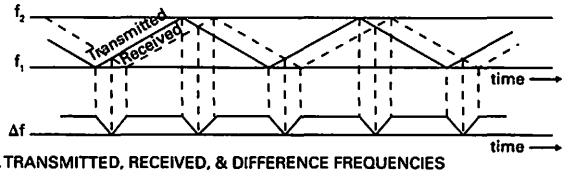
A device that measures frequency can be calibrated to indicate range (distance to the object). Combining equation 4-5 and the basic range equation for two-way transmission, range is calculated as:

$$R = \frac{cT\Delta f}{2(f_2 - f_1)} \qquad (4\text{-}6)$$

FMCW systems are often used for radar altimeters, or in warhead radar proximity fuzes. A frequency-modulated (FM) radar altimeter found on board some aircraft and cruise missiles commonly uses this technique. Its function is to measure the absolute altitude of the aircraft—that is, to measure the range of the earth from the aircraft. A simple block diagram of a typical radar altimeter is shown in figure 4-6a. Figure 4-6b shows the transmitted-received sawtooth form of a modulated signal.



A. BLOCK DIAGRAM



B. TRANSMITTED, RECEIVED, & DIFFERENCE FREQUENCIES

FIGURE 4-6 *Radar altimeter*

In this example, frequency of the transmitted output is swept from f_1 to f_2 and back again, both excursions being executed at a linear rate. The reflected wave from a point on the earth's surface is identical to the transmitted wave except for a time delay, and hence frequency difference, which depends on altitude. Inside the radar, transmitted and received frequencies beat together to produce a difference frequency, Δf. Several possibilities exist for displaying the output data. One such system produces a direct-output voltage proportional to frequency difference, which is read directly on a radar altimeter.

Another way to construct an FMCW system is to compare the phase difference between transmitted and received signals after they have been demodulated to recover sweep information. This system does not need to discriminate the negative values of Δf. In either case, however, the maximum unambiguous range will still be determined by the sweep period of f_1 to f_2 (T), namely[iii]:

$$R_{unamb} = \frac{cT}{2} \qquad (4\text{-}7)$$

FMCW systems do not have a minimum range like a pulsed system does. However, they are not suitable for long-range detection, because the continuous power level they transmit is considerably lower than the peak power of a pulsed system. The proximity of the two antennas also limits the peak power output. Recall that peak and average power in a pulse system are related by the duty cycle $P_{avg} = DC \times P_{peak}$. For a continuous-wave system, the duty cycle is one. Alternatively, peak power is the same as average power. For comparison, in pulsed systems the peak power is many times greater than the average.[iv]

Combination Radar Systems

A factor that sometimes seriously degrades the performance of search radars and pulse radars in general is clutter. Returns from stationary or low-velocity targets such as trees, waves, buildings, terrain features, and so on, produce clutter. Clutter is defined as "unwanted echoes, typically from the ground, sea, rain, or other precipitation." The effect of stationary targets is to clutter the display and make detection of ships and aircraft more difficult. A system that rejects clutter but accepts moving targets is often highly desirable.

As an example, consider a target aircraft flying close to the ground in order to avoid detection. Returns from

the target and any nearby terrain features would arrive simultaneously. If conventional radar were in use, the target return would be masked by the return from the larger terrain features. A radar system sensitive to moving targets, however, would reject the ground return and thereby highlight the moving target.

The purpose of this section is to examine methods that allow rejection of stationary targets by a pulse radar. The fundamental principle of operation depends upon the fact that echoes from a stationary target will return to the radar site after a fixed delay. On the other hand, the delay associated with a moving target will vary from pulse to pulse by increasing or decreasing, depending upon whether the target's range is opening or closing.

Velocity-discriminating radar sets use a modulated pulse system to sense target motion and may be designated either as pulse-Doppler (PD) radar or as moving target indicator (MTI) radar. Both types of radar operate under the same physical principles, and the distinction is generally based upon PRF. Radars characterized as PD operate at higher PRFs, typically greater than 100 KHz, and provide better velocity discrimination and clutter rejection. Radars characterized as MTI operate at lower PRFs, less than 4 KHz, and provide more accurate range resolution.[v]

Velocity Discrimination

If a basic CW radar system were modified to pulse the transmitter energy instead of transmit the energy continuously, a simple system for detecting moving targets would be the result. A diagram of such a system is shown in figure 4-7.

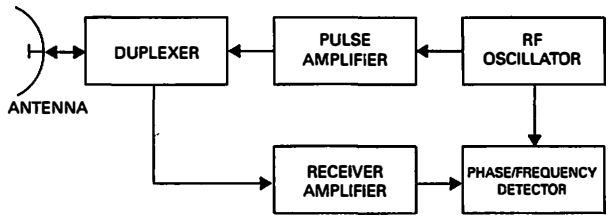


FIGURE 4-7 *Elementary pulse-modulated Doppler system*

Essentially, the system is a pulse system, with one important difference: the receiver is sensitive to frequency shifts within the return pulse. Just as with CW radar, there would be two frequencies to compare, the frequency of the transmitted pulse from the oscillator and the frequency of the received pulse. The important distinction here is that this is a pulsed radar and does not transmit continuously. Therefore the comparison is between the transmitted carrier frequency and the return frequency of each pulse and is used to detect Doppler shift.

As with CW radars, Doppler shift is proportional to radial velocity or v_{LOS}:

$$v_{LOS} = \frac{\lambda f_D}{2} \quad \text{or} \quad f_D = \frac{2v_{LOS}}{\lambda} \tag{4-8}$$

Pulse-Doppler Radar Mode

To avoid velocity ambiguities, pulse-Doppler radars use a higher PRF. By routing the Doppler shift through a frequency-filtering device, any unwanted nonmoving returns are selectively eliminated. Such a device is called a bandpass filter. It allows passage of only those signals that possess frequency characteristics within the narrow band for which the filter was designed. Filters are sensitive to velocity changes on the order of 3 meters/sec. Signals from stationary targets can therefore be blocked from reaching the indicator.

A series of bandpass filters can be used to provide a means of distinguishing between targets of different velocity characteristics, as shown in figure 4-8. In airborne applications, stationary targets would appear to be closing with a relative velocity equal to the velocity of the transmitting aircraft. In order to display only moving targets, it is a fairly simple matter to selectively "turn off" the appropriate bandpass filter that corresponds to this velocity.

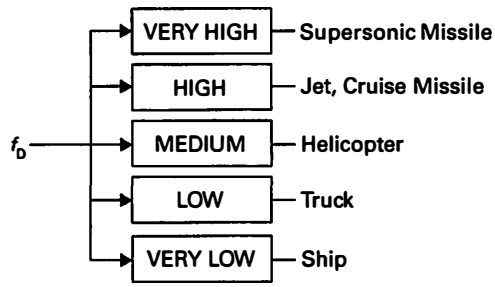


FIGURE 4-8 *Doppler frequency gates for separation of targets*

Radar systems that derive target velocity information in this manner are designated as pulse-Doppler radars. These radars possess advantages over CW systems. One advantage is that since they are pulse modulated, they can measure range as well as velocity. Another is that a pulse-Doppler system requires only one antenna instead of two, and routinely the power of transmission is greater.

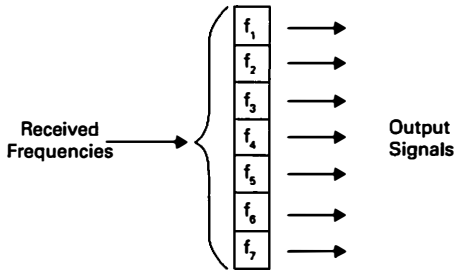


FIGURE 4-9 *Doppler filter bank*

Doppler filter bank In actual practice, pulse-Doppler radars use several banks of bandpass filters whose outputs are linked to an indicator. The radar receives signals from many sources simultaneously, which are then sorted out on the basis of Doppler frequency in a bank of Doppler filters.

Each filter in the bank is designed to pass a narrow band of frequencies, f_i; see figure 4-9. Ideally, by neglecting filter sidelobes, an output is produced only when the frequency of the received signal falls within the specific range of frequencies called the passband of the filter. The closer the received signal is to the center frequency of the passband, the greater the output, as shown in figure 4-10. Each filter in the bank is then tuned to a particular center frequency such that the passbands of each are allowed to overlap; see figure 4-11.

Filter bandwidth A filter achieves its selectivity by integrating the signals applied to it over a period of time;

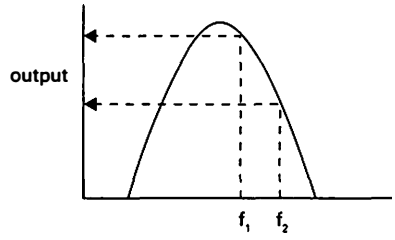


FIGURE 4-10 *The passband of one Doppler filter; f_1 is closer to the center frequency of the filter and produces more output than f_2*

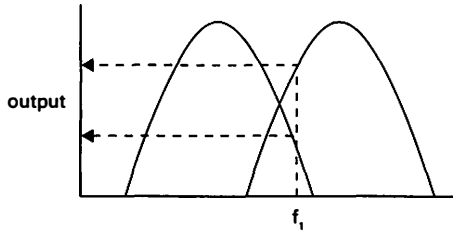


FIGURE 4-11 *With this configuration, as Doppler frequency changes due to motion of the target, output is produced primarily from one filter and then the next and so on.*

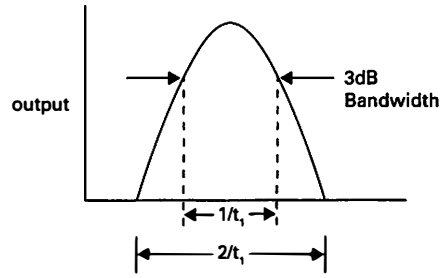


FIGURE 4-12 *Graphical representation of 3dB bandwidth*

this is called the filter bandwidth and defines the minimal resolvable Doppler shift. The width of the band of frequencies passed by the filter then depends primarily upon the length of the integration time t_{int} (on the order of milliseconds) that is applied to the return signal; see figure 4-12. It can be shown for passband filters that the optimum 3dB bandwidth is approximately equal to $1/t_{int}$ or:

$$BW_{3dB} \cong \frac{1}{t_{int}} \tag{4-9}$$

Though the mathematics of Doppler filtering is beyond the scope of this text, in general the filtering is carried out by sampling the return data using the discrete Fourier transform (DFT) method, usually implemented in the form of the fast Fourier transform (FFT) method.

The most common application is to color-code the return information on the PPI display. Doppler shift is sorted into categories—for example, positive, zero, and negative—which are then associated with colors. Target relative velocity and bearing are thereby displayed in the same manner that range and bearing are displayed for pulse-echo-type radar sets.

Pulse-Doppler radar systems are used in numerous military applications. They are also the standard weather radar used throughout the country. Figure 4-13 illustrates

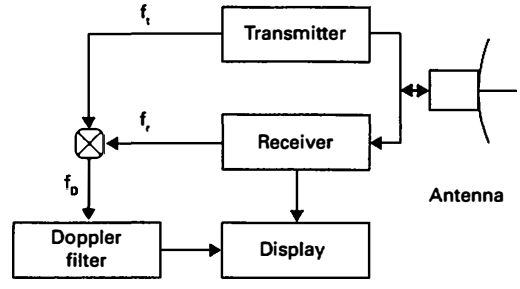


FIGURE 4-13 *Pulsed Doppler radar system*

this type of radar system. Pulse-Doppler radars can detect and graphically display information about the relative motion of winds inside storm cells and have proved useful in detecting tornadoes. For example, a Doppler velocity display of a tornado will show the two colors that correspond to opposite directions of motion side by side.[vi]

MTI Mode

Currently, radars operating in the MTI mode are generally operated at low PRFs, typically lower than 4 KHz. The low PRF makes it better suited to track slower-moving targets such as ground vehicles, ships, or perhaps even a submarine's periscope. However, using the low PRF causes velocity to be ambiguous.

An older technique for MTI used a delay line canceller and phase comparator to detect motion, but like current MTI radars, still could not measure velocity accurately. Figure 4-14 illustrates this type of radar system.

Successive return pulses are sent from a receiver to a phase detector, where they are compared to a sample of the transmitted signal. Output of the phase detector for both fixed and moving targets is shown in figure 4-15.

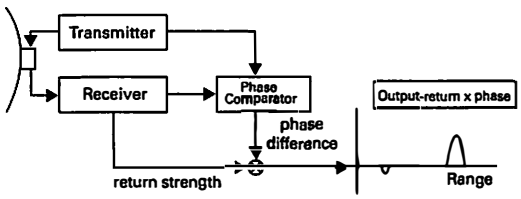


FIGURE 4-14 *MTI phase comparator radar system*

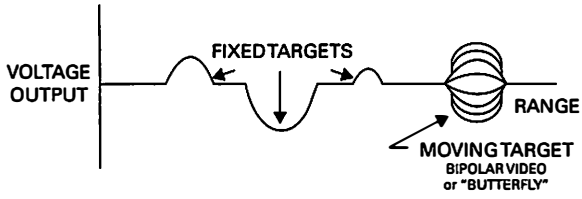


FIGURE 4-15 *An "A" scope presentation of phase detector output for fixed and moving targets*

When range to a target is changing, the phase comparison output will be varying between its extreme values, as well as moving in range. One full cycle of phase shift is completed as the range changes by one-half wavelength of the radar carrier frequency. This is because the radar signal travels both to and from the target, so the total distance traveled by the radar pulse changes by a factor of 2. For a typical radar wavelength of 3 centimeters, it is clear that the phase comparison output will rapidly vary for targets whose range is changing.[vii]

This system determines motion by delaying the output of the phase comparator for an interval equal to 1 PRT and then subtracting it from the next pulse. The signals that appear the same for all pulses are those from stationary targets and are therefore cancelled. Signals from moving targets are not cancelled and may, in fact, be enhanced. The outputs of the phase detector and the delay line canceller—see figure 4-16—are compared in figure 4-17.

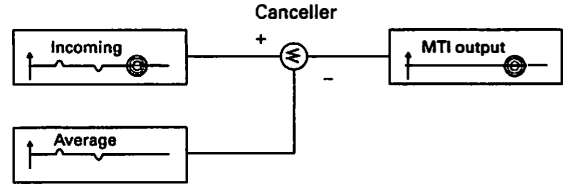


FIGURE 4-16 *Cancellation circuit of MTI processor*

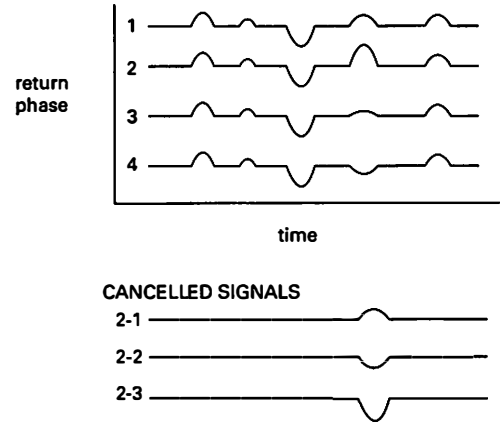


FIGURE 4-17 *Cancellation of echoes*

The fact that stationary targets have a fixed value of phase difference can be exploited when used with a non-moving transmitter to remove them from the display. A cancellation circuit accomplishes this; see figure 4-17. The MTI processor takes a sample from the phase comparison output and averages it over a few cycles. Moving targets will average to zero, while stationary targets will have nonzero averages. The average signal is then subtracted from the output before it is displayed, thereby cancelling out stationary targets.[viii]

Note: Stationary targets are those returns that are not changing in range. For moving transmitters, of course, returns from fixed objects on the ground will be changing in range and therefore displayed. MTI systems for moving transmitters must provide a modified input to the phase comparator, which includes the phase advance associated with the motion of the transmitter.

MTI radars are also useful for clutter rejection, but are usually better suited to search functions. Because of their limitations in unambiguous velocity measurement, MTI radars are usually classified as motion or movement detectors. For example, an aircraft might use an MTI radar mode to search for an inbound aircraft flying at low level. Although the MTI would accomplish some clutter rejection, it would likely have some velocity ambiguities. To overcome these ambiguities, the radar would need to be switched to a higher PRF, PD mode.

Velocity Limitations

In the previous sections concerning pulse-Doppler and MTI modes, velocity ambiguities were mentioned. These ambiguities are called blind speeds and maximum unambiguous speeds. Another limitation is that Doppler radars are restricted to some absolute lower value of velocity that can be measured due to the bandwidth.

Speed limitations of MTI and pulse-Doppler radar systems exist because they are unable to measure any velocity at discrete values of the PRF and are unable to unambiguously measure velocities beyond a certain value, because of their pulsed nature. If a target is moving at a particular velocity such that the phase change from consecutive pulses is 360°, or an exact multiple of 360°, then the target return will not fluctuate. This occurs when the Doppler frequency, f_D, is exactly equal to the PRF or a multiple thereof. This can be shown by looking at the distance a target travels, ΔR, for a given target velocity in the line-of-sight, v_{LOS}, for two consecutive radar pulses. This distance traveled can be found from:

$$\Delta R = 2(v_{LOS})(\text{PRT}) \tag{4-10}$$

When $\Delta R = \lambda/2$ or $n(\lambda/2)$, there is no phase change or apparent frequency change observed at the radar receiver. Mathematically this statement is written:

$$v_{LOS}\text{PRT} = n(\lambda/2)$$

or

$$v_{LOS} = \frac{n\lambda}{2\text{PRT}} = \frac{n\text{PRF}\lambda}{2} \tag{4-11}$$

These target speeds are called blind speeds, since a delay line canceller would detect no pulse-to-pulse changes in phase. Further, using equation 4–3 and assuming a stationary source:

$$f_{D\text{–blind}} = n\text{PRF} \tag{4-12}$$

EXAMPLE 4-1: Calculate the blind speeds for a radar with the following specifications:

$$\text{wavelength} = 0.1 \text{ m} \qquad \text{PRF} = 200 \text{ Hz}$$

SOLUTION:

$$v_{LOS} = \frac{n}{2} \times \frac{200}{\text{sec}} \times 0.1 \text{ meter} = 10n \frac{\text{meters}}{\text{sec}}$$

Therefore, the radar would be blind in terms of velocity to all targets having relative radial velocities of $n = 10, 20, 30 \ldots$ meters/sec. Converting to knots, the values of target speeds extending from 0 to 600 knots would contain 31 blind speeds, each occurring at integer multiples of 19.3 knots.

For search radars, this multiplicity of blind speeds is of little consequence since they fall within a narrow range and are sharply defined. Since it is unlikely that a target on any normal trajectory would approach the site for more than a few seconds at one of the blind speeds, the probability of detection of such targets would remain quite high. If, however, it is deemed necessary to minimize the blind speeds within a specified range of values, it is possible to do this by varying the pulse repetition rate of the radar. One method of accomplishing this is to have every second pulse delayed by a small time interval. The result is a staggered PRF as shown in figure 4-18.

Referring to figure 4-18, notice that the shorter interval, t_1, produces a blind speed that is higher than that associated with the longer interval $t_1 + \Delta t$. It is possible to choose t_1 and Δt such that a blind speed due to one of these intervals will usually not be blind for the other.

Most Doppler systems avoid the problem of blind speeds by making the PRF high enough so that the first blind speed is higher than the highest target velocity likely to be encountered.

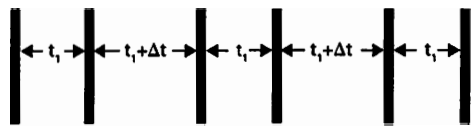


FIGURE 4-18 *Staggered pulse repetition frequency*

EXAMPLE 4-2: What PRF is needed if a radar with a wavelength of 0.10 meter is to have no blind speeds under 600 knots (relative radial speed)?

SOLUTION: Convert 600 knots to 304.8 meters/sec.

$$\text{PRF} = \frac{2v_{\text{LOS}}}{\lambda} = \frac{609.6 \text{ meters/sec}}{0.1 \text{ meter}} = 6{,}069 \text{ Hz}$$

Such high pulse repetition rates introduce yet another problem—that of ambiguous range. For example, a pulse-Doppler radar with a PRF of 6,000 Hz would have a maximum unambiguous range of only about 25 kilometers. Range ambiguities can also be resolved through the use of a staggered PRF. For this reason some long-range missile fire control radar's pulse repetition rates continuously change. Another method involves the use of frequency agility. In this method, the frequency of the radar is changed from pulse to pulse so that each echo can be identified with a particular outgoing transmitted pulse.

A similar concept to blind speeds is the maximum unambiguous velocity, v_{unamb}. In pulsed Doppler mode, the radial velocity cannot be measured unambiguously above the first blind speed. This is much akin to maximum unambiguous range. Any speed greater than this speed cannot be unambiguously measured.

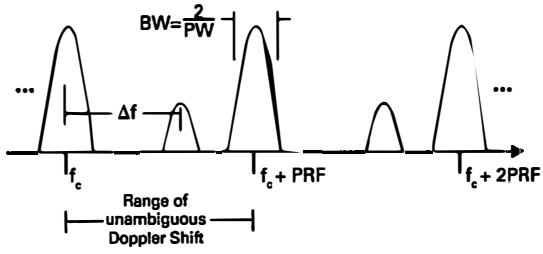


FIGURE 4-19 *Spectrum of pulsed Doppler radar*

The condition of maximum unambiguous velocity occurs when the Doppler shift is equal to the PRF. Consider the spectrum of the pulsed Doppler radar system in figure 4-19. There will be a main lobe centered on the radar carrier frequency. As shown in chapter 3, the bandwidth (BW) of the main lobe is determined by the pulse width (PW), using the approximate relationship:

$$BW \cong \frac{2}{\text{PW}} \tag{4-13}$$

Because the cycle is repeated at a frequency equal to the PRF, there are additional copies of the main lobe at multiple intervals of the PRF on either side of the carrier frequency. All of the information from the return will be repeated at intervals of PRF, including the Doppler-shifted return.[ix]

From this, the maximum unambiguous speed is given by[x]:

$$v_{\text{unamb}} = \frac{\lambda \text{PRF}}{2} \tag{4-14}$$

Therefore to achieve greater maximum unambiguous velocities, higher PRFs are needed. This is in contrast to maximum unambiguous range, R_{unamb}, which requires lower PRFs for longer maximum unambiguous range.

Doppler radars also have an absolute limit to the minimum velocity that can be measured. From the spectrum in figure 4-19, it is also apparent that the Doppler shift must be larger than one-half the bandwidth of the main lobe in order to be detectable:

$$f_D = \frac{1}{2}\text{BW} \tag{4-15}$$

This defines a minimum detectable speed in the line-of-sight when one substitutes equation 4–14 into 4-3 and solves for minimum velocity in the LOS:

$$v_{\text{min}} = \frac{\text{BW}\lambda}{4}$$

or in terms of PW: (4-16)

$$v_{\text{min}} = \frac{\lambda}{2\text{PW}}$$

TABLE 4-1 R_{unamb} *and* v_{unamb} *for a 10 GHz radar*

PRF (KHz)	R_{unamb} (km)	v_{unamb} (m/s)
0.1	1,500	1.5
0.5	300	7.5
1	150	15
4	37.5	60
10	15	150
50	3	750
100	1.5	1,500

Table 4-1 illustrates the relationship between PRF, unambiguous range, and unambiguous velocity; notice there are advantages to both ends of the spectrum.[xi]

High-Resolution Radars

Pulse Compression

The amount of energy in a rectangular pulse is given by the product of peak power and pulse width. Thus, for a radar system that is peak power limited, an increase in pulse width would lead to increased detection capability. It is also known from the previous chapter that increased pulse width reduces range resolution. Pulse compression is a signal-processing technique that combines the high energy of a wide pulse to enhance detection capability while maintaining the high range resolution of short pulsed transmissions.

One method of pulse compression uses a transmitted pulse that is frequency modulated over the duration of the pulse; see figure 4-20. By linearly increasing the frequency much akin to a FMCW transmission, one finds a method to further resolve targets with overlapping returns; since each part of the pulse has a unique frequency, the two returns can be completely separated.

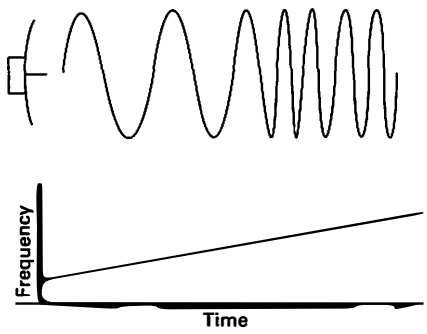


FIGURE 4-20 *Pulse compression using frequency modulation*

When the reflected pulse is received, it is passed through a pulse compression filter in which lower frequencies pass more slowly than higher frequencies. Thus the frequency variation of the returning pulse results in the trailing edge of the pulse moving through the filter faster than the leading edge. This results in the pulse "piling up" on itself. The output of this filter is a pulse of greater power amplitude and narrower pulse width. This modified pulse is then processed in the normal manner.

Figure 4-21 shows the effect of using pulse compression in signal processing. Notice that energy or area contained

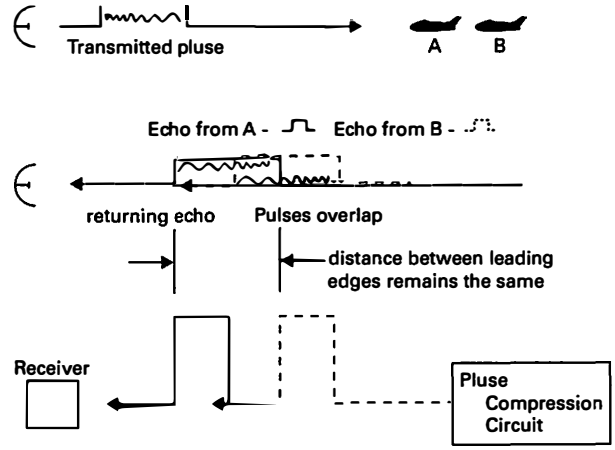


FIGURE 4-21 *Pulse compression*

by the pulses is the same in both cases. In the receiver, pulse compression circuits cause the leading edges of both echoes to be delayed at the same time, with decreasing delay toward the trailing edge of the pulse, resulting in each echo becoming shorter and effectively increasing the amplitude.

When the pulse is frequency modulated in this manner, the process of resolving targets on the basis of frequency is called post-detection pulse compression, or PDPC. The ability of the receiver to improve range resolution over that of conventional systems is called the pulse compression ratio. For example, a pulse compression ratio of 20:1 means that the system range resolution is reduced by 1/20 of the conventional system. Some modern radars have pulse compression ratios as high as 7,000:1. The factor of improvement is given the symbol PCR, which can be used as a number in the range resolution formula, which now becomes:

$$R_{res} = \frac{cPW}{2PCR} \qquad (4\text{-}17)$$

It is important to realize that this process does not improve the minimum range. The full pulse width still applies to the transmission, which requires the duplexer to remain aligned to the transmitter throughout the pulse. Therefore, R_{min} is unaffected.[xii]

Synthetic Aperture Radar (SAR)

While pulse compression can improve range resolution, the angular resolution is another area for improvement.

The angular resolution of a radar beam is a function of the range, R, and the antenna beam width. At a given range, the ability to resolve objects in the cross-range direction known as the cross-range resolution, R_{cross}, is calculated by:

$$R_{cross} = R\theta \tag{4-18}$$

where θ is the beamwidth expressed in radians.

This is merely the arc length swept out by the angle θ at radius R. It is also the width of the radar beam at range R.

EXAMPLE 4-3: A 1° beamwidth at 10 kilometers would have what resolution in cross range?

SOLUTION:

$$R_{cross} = 10{,}000 \text{ m} \times \frac{1\text{deg}}{57.3 \text{ rad/deg}}$$
$$= 174 \text{ meters}$$

This is a very poor resolution for identifying targets; therefore a very narrow beamwidth is needed. Beamwidth θ_{-3dB} is given by equation 3-7:

$$\theta_{-3dB} = \frac{k\lambda}{l} \tag{4-19}$$

To improve this, the wavelength could be changed to something smaller; however, recall that higher frequencies experience greater atmospheric attenuation, which then limits range. Therefore to achieve a smaller beamwidth, the antenna size l must be made larger.

EXAMPLE 4-4: If a 3 Ghz radar ($\lambda = 0.1$ meter) is used and a 1 meter resolution is desired at 100 kilometers, how large does one side of a square antenna need to be? From equation 4-18:

$$\theta = \frac{R_{cross}}{R} = \frac{1 \text{ m}}{100{,}000 \text{ m}} = 10 \times 10^{-6} \text{ radian}$$

SOLUTION: Using beamwidth equation 4-19 and solving for length, l:

$$l = \frac{k\lambda}{\theta} = \frac{1.02(0.1 \text{ m})}{10 \times 10^{-6} \text{ radian}}$$
$$= 10 \times 10^{3} \text{ meters or 10 kilometers long}$$

Clearly it is not feasible to build an antenna this large.

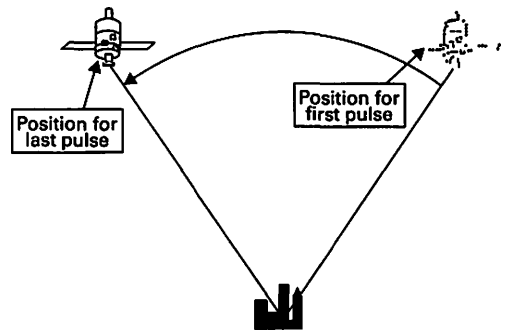


FIGURE 4-22 *Synthetic aperture*

Synthetic aperture radar (SAR) is the solution to this problem because it uses the motion of the transmitter/receiver to generate a large effective aperture. To accomplish this, the system stores several returns taken while the antenna is moving and then reconstructs the signals as if they had all returned simultaneously.

What the system is actually doing is measuring the Doppler shift of various segments of the scene. As the transmitter moves, the target area will appear to rotate. As with CW radars, this is a relative motion problem. If the transmitter/receiver moves a total distance S during the period of data collection and several return pulses are stored, then the effective aperture upon reconstruction is also S; see figure 4-22.[xiii]

This large synthetic aperture creates a very narrow beamwidth. By substituting the synthetic aperture (S) for the physical antenna aperture (l) in equation 4-19, and using a k value for synthetic apertures of 0.5, the cross-range resolution is predicted by:

$$R_{cross} \approx \frac{R\lambda}{2S}(\text{SAR}) \tag{4-20}$$

where:

R = target range
S = distance traveled by the transmitter/receiver during data collection
λ = wavelength of the radar

One application of SAR is in satellite radar systems. Because satellites travel at high velocity, accuracy of these systems can be made very high. Furthermore, if the target is fixed in location, the period for data collection can be made very long without introducing significant errors. Therefore satellite SAR is used for the imaging of fixed objects like terrain, cities, military bases, and so on.[xiv]

SAR-type radars are, however, increasingly found on tactical and reconnaissance aircraft as processing and navigation capabilities are enhanced.

Inverse Synthetic Aperture Radar (ISAR)

The ISAR signal-processing technique is very similar to that of SAR. Whereas SAR uses the motion of the transmitter, ISAR uses the motion of the target. The motion of the target causes a Doppler shift proportional to the angular velocity of the target's motion around a pivot point.

Consider a boat rocking with an angle ψ in the line-of-sight (LOS) of the radar. The rate of rotation is $\Delta\psi/\Delta t$, and the Doppler shift at the top of the mast is given by applying equation 4-3, with $v_{LOS} = (\Delta\psi/\Delta t)\ L$; f_D then becomes:

$$f_D = \frac{2\frac{\Delta\psi}{\Delta t} L}{\lambda} \qquad (4\text{-}21)$$

Dividing the length of the mast into small segments, ΔL, the Doppler shift of any one segment Δf_D is given by:

$$\Delta f_D = \frac{2\frac{\Delta\psi}{\Delta t}\Delta L}{\lambda} \qquad (4\text{-}22)$$

From the geometry of the problem, the mast cannot be divided into segments smaller than the cross-range resolution, R_{cross}. Therefore, ΔL can be replaced by R_{cross}. Rearranging equation 4-22 and solving for R_{cross} gives:

$$R_{cross} = \frac{\lambda}{2\frac{\Delta\psi}{\Delta t}}\Delta f_D \qquad (4\text{-}23)$$

A radar system's ability to resolve differences in Doppler is dependent upon the sampling rate of the incoming signal, called signal integration time, t_{int}. Because of the filtering method used, the system is unable to resolve Δf_D further than $1/t_{int}$. Consequently, equation 4-23 becomes:

$$R_{cross} = \frac{\lambda}{2\frac{\Delta\psi}{\Delta t} t_{int}} \qquad (4\text{-}24)$$

This result shows that R_{cross} is dependent upon the angular rotation rate in the LOS and the Doppler frequency resolution Δf_D of the radar system and is totally independent of range.

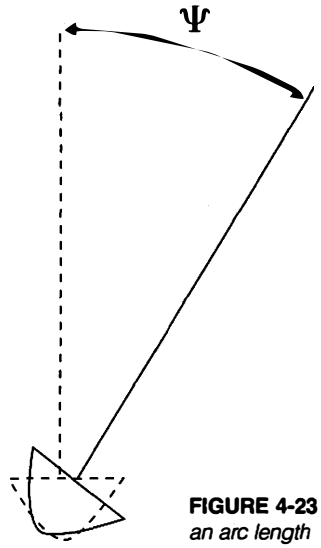


FIGURE 4-23 *Sailboat mast moving through an arc length*

An additional feature of an ISAR radar is that the mast of a sailboat in our example, when examined by a radar, will have a different Doppler shift at the top of the mast than at the deckhouse. The shift will continuously change along the mast's length as the boat rolls, see figure 4-23. By plotting the Doppler shift for each segment versus range, an image of the vessel is given in figure 4-24. This image will slowly shrink and become inverted as the rocking motion causes the Doppler shift to change from positive to negative continuously.

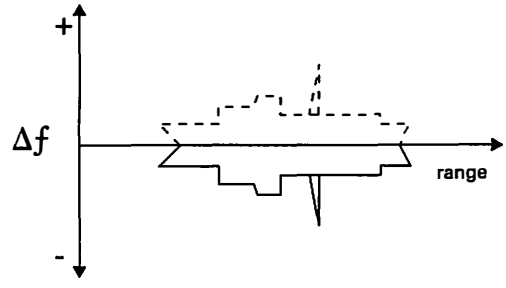


FIGURE 4-24 *Doppler shift versus range showing positive and negative images*

ISAR systems are typically used for long-range imaging and identification of possible targets. ISAR may not be able to identify specific targets but can certainly identify warship/nonwarship and can sometimes identify class depending on aspect angle. ISAR platforms may be fixed or moving, the best targets for ISAR are ships, which move periodically in the sea state.[xv]

Electronic Scanning and the Phased Array

The need for increased target handling capacity is limited by the requirement to position the radar antenna mechanically. Mechanical scanning methods are inherently slow and require large amounts of power in order to respond rapidly enough to handle large numbers of high-speed maneuvering targets. Additionally, mechanically scanning systems have antenna inertia and inflexibility in beam positioning that slow reaction times. With electronic scanning, radar beams are positioned almost instantaneously and completely without the inertia, time lags, and vibrations of mechanical systems. In an era in which the reaction time to threats must be very fast, electronic scanning provides an enormous advantage. The specific benefits of electronic scanning include:

1. Increased data rates and reduction of system reaction time.
2. Virtually instantaneous positioning of the radar beam anywhere within a set sector; beam position can be changed in a matter of microseconds.
3. Elimination of mechanical errors and failures associated with mechanically scanned antennas.
4. Vastly increased flexibility of the radar, facilitating multimode operation, automatic multitarget tracking, highly directional transmission of missile guidance and control orders, and interceptor and general air traffic control from one radar at virtually the same time.

Principles of Operation

The fundamental principles underlying the concept of electronic beam steering are derived from electromagnetic radiation theory, employing constructive and destructive interference. These principles can be stated as follows:

> The electromagnetic energy received at a point in space from two or more closely spaced radiating elements is a maximum when the energy from each radiating element arrives at the point in phase.

To see an illustration of this principle, consider figure 4-25. All elements are radiating in phase, and the resultant wave front is perpendicular to the axis of the element array. Figure 4-26 and subsequent diagrams show only a limited number of radiating elements. In actual radar antenna design, several thousand elements could be used to obtain a high-gain antenna with a beamwidth of less than 2°.

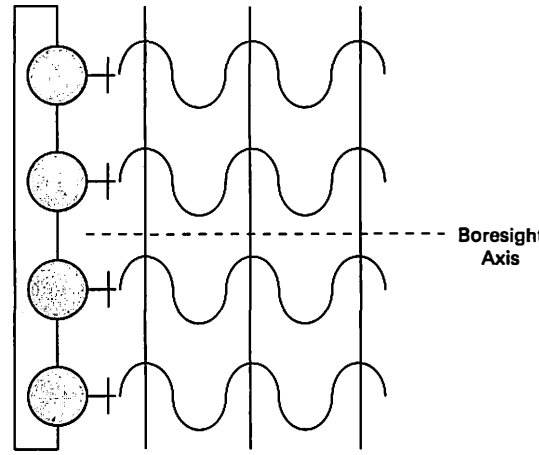


FIGURE 4-25 *Wave front resulting from elements radiating in phase. Maximum energy is concentrated along the array boresight.*

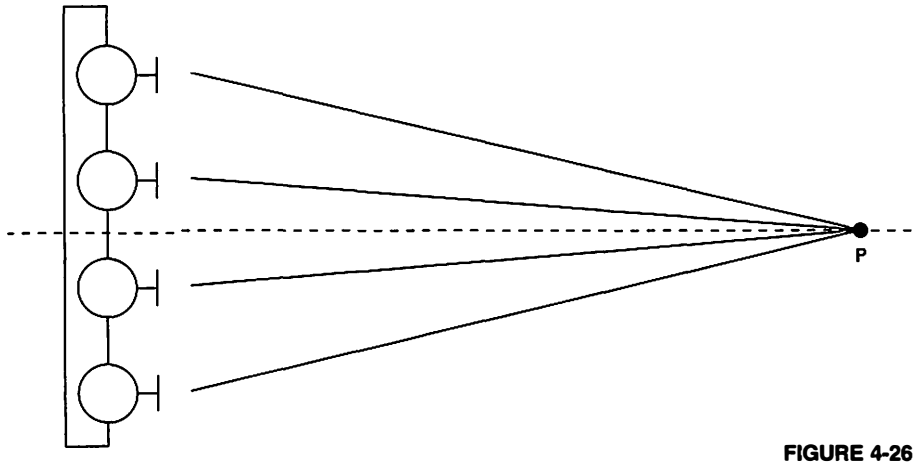


FIGURE 4-26 *Array with elements radiating in phase*

The wave fronts remain perpendicular to the boresight axis and are considered to arrive at a point target in space at the same time. As illustrated in figure 4-26, the path lengths from the elements to point P equalize as P approaches infinity. Thus, in situations where the target range is very large compared to the distance between elements, paths from the elements to point P are almost parallel. Under these conditions, energy will arrive at point P with the same phase relationship that existed at the array.

To achieve beam positioning off the boresight axis, it is necessary to radiate the antenna elements out of phase with one another. Figure 4-27a depicts the phase shift necessary to create constructive interference along a line joining an arbitrary point P with the center of the array. In order to achieve this constructive interference at point P, the energy arriving from all radiating sources must be in phase and arrive at the same time. Energy from element e_1 must travel a path length r_1, while energy from element e_2 must travel the longer path length r_2.

The electric field magnitudes from the two elements at point P are given by equation 2-12, namely:

$$E_1 = E_0 \sin\left[\frac{2\pi}{\lambda}(r_1 - ct) + \phi_1\right]$$

and (4-25)

$$E_2 = E_0 \sin\left[\frac{2\pi}{\lambda}(r_2 - ct) + \phi_2\right]$$

Since these two electric fields must be in phase at point P for constructive interference to occur, the arguments of the sine functions must be equal.

$$\frac{2\pi}{\lambda}(r_1 - ct) + \phi_1 = \frac{2\pi}{\lambda}(r_2 - ct) + \phi_2$$
$$\phi_1 - \phi_2 = \frac{2\pi}{\lambda}(r_2 - ct) - \frac{2\pi}{\lambda}(r_1 - ct) \quad (4\text{-}26)$$

thus

$$\phi_1 - \phi_2 = \Delta\phi = \frac{2\pi}{\lambda}(r_2 - r_1) \quad (4\text{-}27)$$

The path length difference, $r_2 - r_1$, approaches the value $d \sin \varepsilon$ as the distance R in figure 4-27a increases. In figure 4-27b, where point P has been moved an infinite distance from the source, the paths from the two sources have become parallel, and the path length difference is exactly $d \sin \varepsilon$. For points at a distance R less than infinity, $d \sin \varepsilon$ is still a good approximation for the path length difference, $r_2 - r_1$,

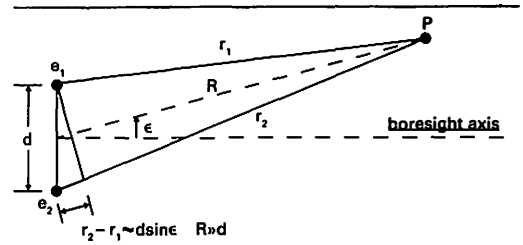


FIGURE 4-27a *Radiation path length difference for energy arriving at point P from two sources*

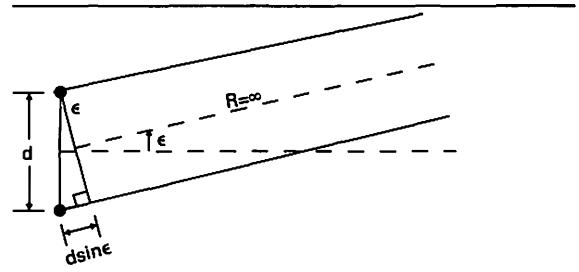


FIGURE 4-27b *Radiation path length difference for energy arriving at a point an infinite distance from two sources*

as long as R is large compared to the element spacing, d. Thus, applying this distance approximation yields:

$$r_2 - r_1 \cong d \sin\varepsilon \quad \text{if } R \gg d \quad (4\text{-}28)$$

For practical radar applications, d (the distance between elements) is on the order of a few centimeters wheras R is on the order of kilometers, a difference of several orders of magnitude; therefore, the distance approximation is valid for all radar applications in the far radar field (ranges greater than 1 kilometer).

Applying the distance approximation to the expression already obtained for the required phase difference between two radiating elements yields:

$$\Delta\phi = \frac{2\pi d}{\lambda}\sin\varepsilon \quad (4\text{-}29)$$

where:

Δ = phase shift between adjacent elements expressed in radians
λ = free-space wavelength in meters
d = linear distance between radiating elements in meters
ε = desired angular offset in degrees

Methods of Beam Steering

The previous discussion addressed the theory required to compute the relative phase shift between adjacent

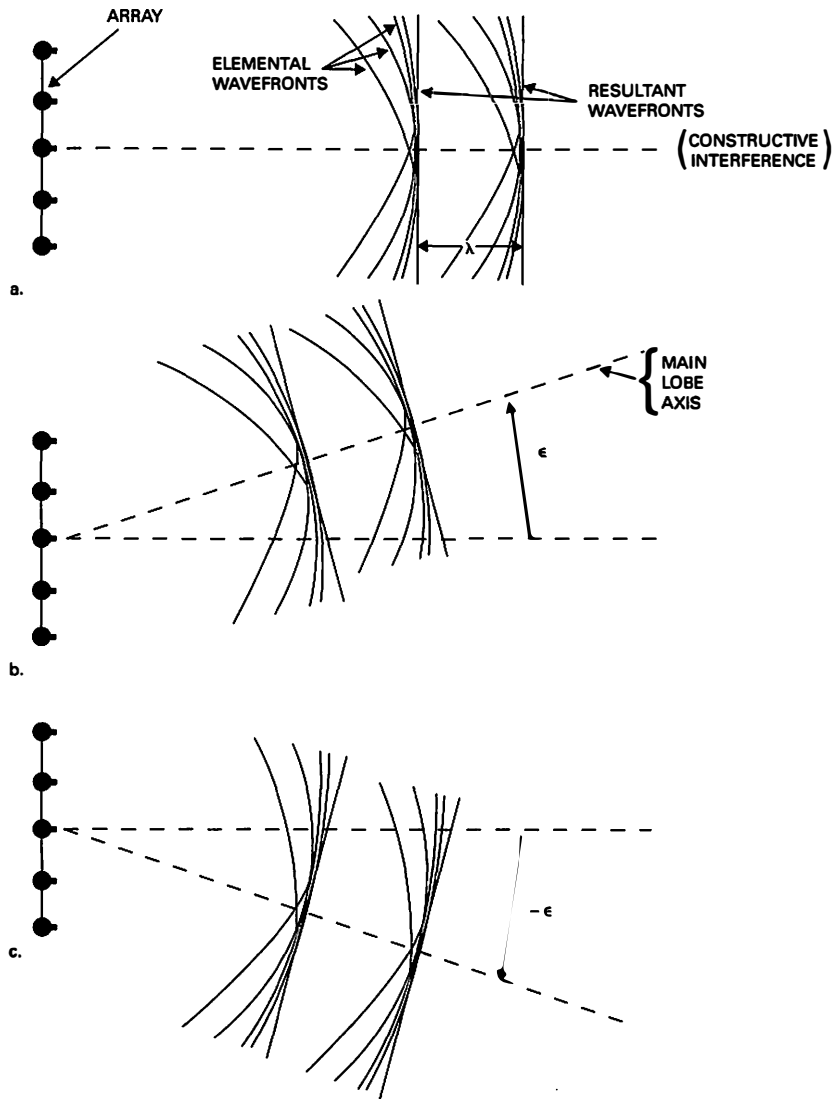


FIGURE 4-28 *(a) Beam positioning along the boresight axis; all elements are radiated in phase. (b) Beam positioning above the boresight axis; elements lag in phase by Δϕ with the uppermost element receiving the greatest phase shift. (c) Beam positioning below the boresight axis; elements lag in phase by Δϕ with the lowermost element receiving the greatest phase shift.*

radiating elements in order to position the beam of an array-type antenna to a specific angle off the antenna boresight. In practice, there are three methods of accomplishing this phase difference: by using time delays, by changing frequencies, and by changing the phase of each element.

Time-Delay Scanning

Employment of time delay as a means of achieving the desired phase relationships between elements allows greater flexibility in frequency utilization than other methods. However, in practice the use of coaxial delay lines or other means of timing at high power levels is impractical due to increased cost, complexity, and weight.

To accomplish time-delay scanning, variable delay networks are inserted in front of each radiating element; see figure 4-29. By proper choice of these time delays, the required effective phase shift can be applied to each element, The time delay between adjacent elements required to scan the beam to an elevation angle, ε, is given by:

$$t = \frac{d}{c} \sin \varepsilon \qquad (4\text{-}30)$$

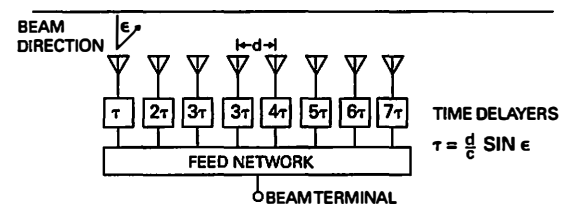


FIGURE 4-29 *Time-delay scanning*

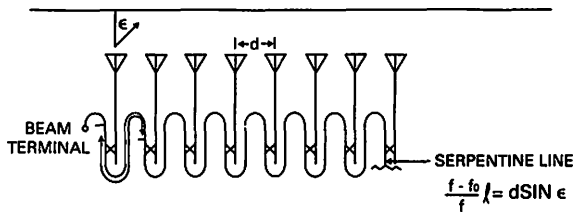


FIGURE 4-30 *Frequency scanning*

Frequency Scanning

One of the simpler methods of phased-array radar implementation is frequency scanning. This method is also relatively inexpensive. Figure 4-30 shows the schematic arrangement of elements when frequency scanning is used to position a beam in either azimuth or elevation. The length of the serpentine wave guide line (l) is chosen such that for some center frequency f_0, the length of signal travel between elements is an integral number of wavelengths, or:

$$l = n\lambda_0 \quad (n = \text{any integer greater than } 0) \tag{4-31}$$

where:

λ_0 = wavelength in the serpentine line at frequency f_0

Thus, when the excitation frequency is f_0, the serpentine line will cause all elements to radiate in phase, and the beam will be formed along the boresight axis. If the excitation frequency is increased or decreased to f from the center frequency, f_0, the line length, l, will no longer represent an integer number of wavelengths.

As the excitation energy travels along the serpentine line, it will reach each successive radiating element with a uniformly increasing positive or negative phase shift. This results in the beam being deflected by an angle from the boresight axis. Thus, by varying the radar transmitter frequency about some base frequency, the beam can be positioned in one axis as shown by:

$$\varepsilon = \sin^{-1}\left[\frac{l(f - f_0)}{f_d}\right] \tag{4-32}$$

In figure 4-31 the frequency-scanned array has been simplified to a two-element system with the boresight axis

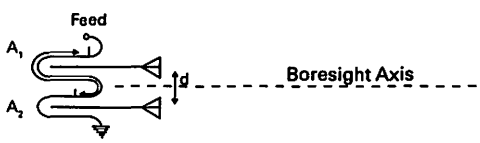


FIGURE 4-31 *Two-element frequency-scanned array*

normal to the plane of the elements. The feed is folded into a serpentine form to allow close element spacing while maintaining the required line length (l) between elements.

If RF energy at 5,000 MHz is fed at the top of the array, and the elements are separated by distance d equal to (0.03 meter). At time $t = 0$, energy enters the serpentine feed line, and antenna A_1 radiates immediately starting at zero phase. Since the period of the wave form is $T = 1/f$, then T = 1/5,000 MHz or 200p sec. Therefore, it takes 200p sec for one wavelength to propagate from A_1. If the distance l traveled in the serpentine feed between A_1 and A_2 equals one wavelength or any integer number of wavelength ($l = n$ where $n = 1, 2, 3, \ldots$), then,

$$t = \frac{l}{c} = \frac{n\lambda_0}{c} = \frac{n\left(\frac{c}{f_0}\right)}{c} = \frac{n}{f_0} = nT \tag{4-33}$$

where:

t = elapsed time

T = period of the waveform

Therefore, the energy from A_2 will always be in phase with A_1. The beam formed by the array will be on the boresight axis. Note that this represents a broadside array because the elements are transmitting in phase. If the frequency is changed to 5,500 MHz, the period T becomes 181.81p sec; however, the waveform still takes $t = l/c$ or 200p sec longer to reach A_2 than to reach A_1 when fed into the serpentine line as depicted in figure 4-32.

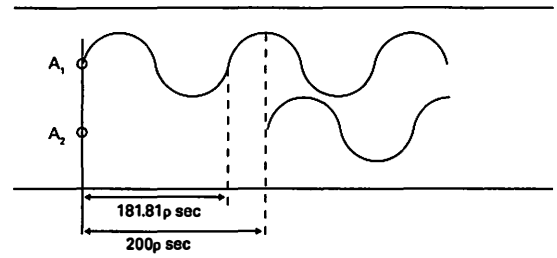


FIGURE 4-32 *Waveforms with delay in phase*

In this case, the wave from A_2 is no longer in phase with A_1. Energy from A_2 lags A_1 in phase by 200p sec–181.81p sec or 18.19p sec. The amount of phase shift is determined by:

$$\frac{t}{T} = \frac{\Delta\phi}{2\pi}$$

$$\frac{18.19\text{p sec}}{181.81\text{p sec}} = \frac{\Delta\phi}{2\pi}$$

$$\Delta\phi = (0.1)(2\pi) = 0.6286 \text{ radian or } 36°$$



FIGURE 4-33 *Beam depression due to phase lag*

Since there is a phase difference, $\Delta\phi$, the beam axis is located using equation 4-29 as follows:

$$\Delta\phi = \frac{2\pi d}{\lambda}\sin\varepsilon$$

$$0.6286 = \frac{2\pi(0.03\text{ m})\sin\varepsilon}{(0.0545\text{ m})}$$

$$\sin\varepsilon = 0.1817$$

$$\varepsilon = 10.47^\circ$$

Since energy from A_2 lags 36° in phase, the beam will be 10.47° below the boresight, as shown in figure 4-33.

In this illustration, the distance from A_1 to a point in space (R_1) is greater than that from A_2 to that point (R_1). The waveforms will arrive in phase at the point because A_2 lags A_1 when the energy is transmitted.

Thus, as frequency is varied, the beam axis will change, and scanning can be accomplished in one axis (either elevation or azimuth). These principles are employed today in the AN/SPS-48 series radars. Variation in frequency tends to make these radars more resistant to jamming than if they were operated at a fixed frequency, and this method also provides a solution to the blind speed problem of MTI systems.

Frequency scanning does impose some limitations, however, in that a large portion of the available frequency band is used for scanning rather than optimizing resolution of targets. Additionally, this imposes the requirement that the receiver bandwidth be extremely wide or that the receiver be capable of shifting the center of a narrower bandwidth to the transmitted frequency. Equation 4-34 gives the relationship between the percentage variation in frequency, $\Delta\%f$, and the scan angle, ε,which is referred to as the wrap-up ratio.

$$\frac{l}{d} = \frac{\sin\varepsilon_{max}}{\Delta\%f} \qquad (4\text{-}34)$$

where

$$\Delta\%f = \left|\frac{100\,(f - f_0)}{f}\right|$$

The wrap-up ratio is the ratio of the sine of the maximum scan angle to the percentage change in frequency required to scan.

Phase Scanning

In a phase-scanned radar system, radiating elements are fed from a radar transmitter through phase-shifting networks or "phasers." This system is shown in figures 4-35, 4-36, and 4-37. The aim of this system is again to position the beam at any arbitrary angle, at any time. In this case, the means of accomplishing the phase shift at each radiating element is simply to shift the phase of the incoming energy to each element. These phasers are adjustable over the range 0 to $\pm 2\pi$ radians. The task of the system is then to compute the phase shift required for each element, and set each phaser to the proper value to accomplish the desired beam offset.

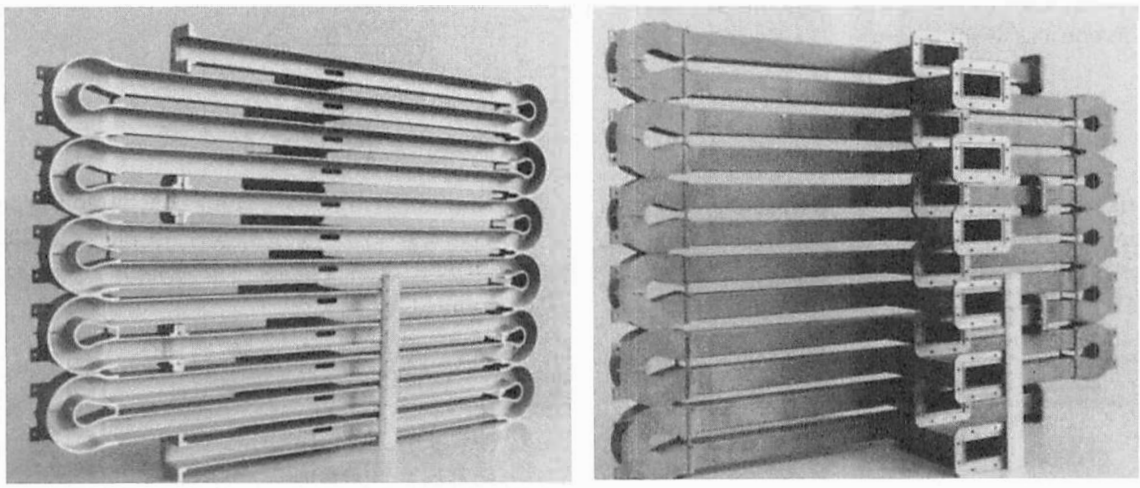


FIGURE 4-34 *Serpentine feed assembly for a frequency-scanning radar, cut away to show the internal detail*

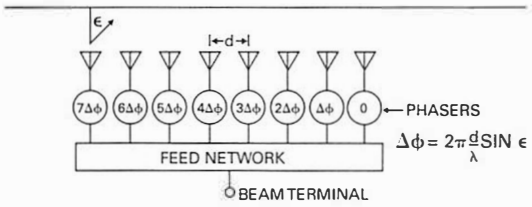


FIGURE 4-35 *Phase scanning*

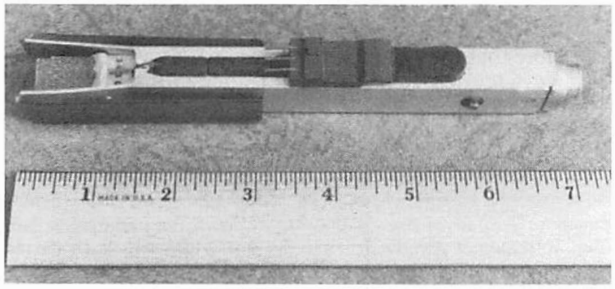


FIGURE 4-36 *A ferrite phase shifter or "phaser"*

While phase scanning is more expensive than frequency scanning, it is much less expensive (in dollars, weight, and power losses) than time-delay scanning. To a first approximation, the bandwidth of a phase-scanning antenna in percent of f_0 is equal to the normal beamwidth in degrees. Thus a 1° beamwidth antenna working at 10 GHz can radiate over a 100 MHz bandwidth (i.e., +50 MHz) without distortion.

Energy Reception

To receive energy transmitted in a steered beam by any of the three scanning methods, the applied frequency, time, or phase relationships are maintained at each element, which has the effect of making the radar sensitive to energy from the direction of transmission. Thus, each of the scanning methods is completely reversible and works equally well in the reception of energy as in transmission.

Unfortunately, some beam positions do not have the same phase shift in the reverse direction. When nonreciprocal phase shifters are used, it is necessary to change the phase shifter setting between transmit and receive to use the same phase shift on receive as was used on transmission.

Computation of Required Phase Relationships

No matter which of the three possible methods of scanning is used in a phased-array system, the objective is a relative phase shift of the energy being radiated by each element in the array. The incremental phase shift required between two adjacent elements is given by equation 4-29. When using this equation, it will be assumed for consistency that $\Delta\phi$ represents the phase lead given to each element with respect to its adjacent element in the direction of the chosen reference element.

Thus, when positioning the beam to the same side of the boresight axis as the reference element, each array element must lead the next element closer to the reference by the same amount, $\Delta\phi > 0$. When positioning the beam on the opposite side of the boresight axis from the reference by the same amount, $\Delta\phi < 0$ (negative or effective phase lag). This convention can be extended to include a sign convention for the azimuth angle α. Choose the reference element as the topmost element and the farthest-right element when looking outward from the antenna along the boresight axis. Also choose an elevation angle, ε,

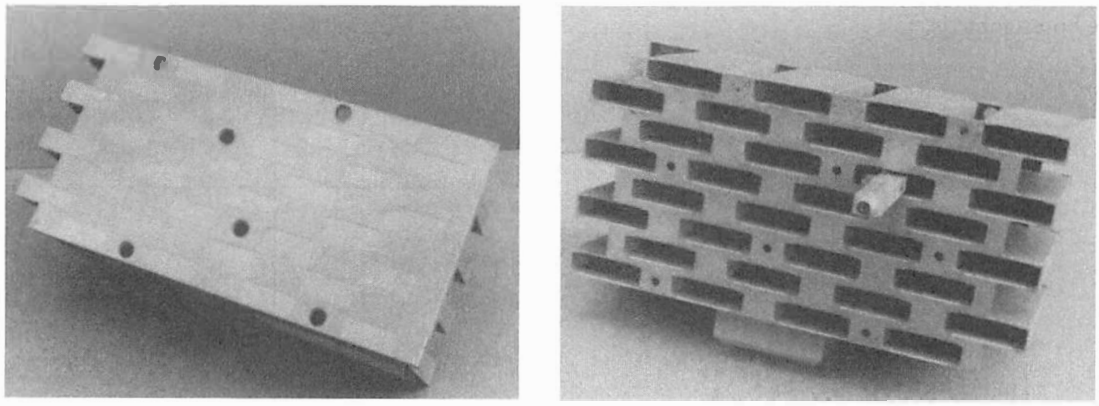


FIGURE 4-37 *A section of a phased-array radar antenna. Note the location of the phase shifter in the back of the assembly.*

as positive above the boresight axis and negative below, and choose azimuth angles α as positive in a clockwise direction (scanning to the right) and negative in a counterclockwise direction. (Do not confuse the above definitions of lead/lag in the spatial domain with the electrical definitions of lead/lag in the radian domain.)

Figure 4-38 illustrates this convention. In order to position the beam above the boresight axis, the angle ε will be positive and thus sin ε will be positive also. This yields a positive $\Delta\phi$ between elements. To determine the phase applied to each element ϕ_e, simply use the relationship:

$$\phi_e = e\Delta\phi, \qquad e = 0, 1, 2, \ldots$$

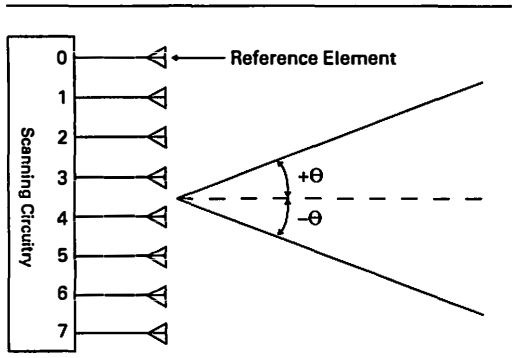


FIGURE 4-38 *Reference element and sign convention for scanning in elevation*

Similarly, to position the beam below the boresight axis requires the use of a negative angle ε. This yields a negative $\Delta\phi$, and equation 4-29 again yields the applied phase of each element.

EXAMPLE 4-5: Phase scanning in elevation. If the reference element has a phase of zero, compute the phase applied to element five when the beam is scanned 40° above the boresight axis.

SOLUTION: Using data supplied in figure 4-39.

$$\text{Find } \lambda: \quad \lambda = \frac{c}{f} = 0.06 \text{ m}$$

$$\text{Find } \Delta\phi: \quad \Delta\phi = \frac{2\pi d}{\lambda} \sin\varepsilon$$

$$\Delta\phi = \frac{2\pi(0.03)}{0.06} \sin 40^\circ$$

$$\Delta\phi = 2.02 \text{ radians}$$

$$\text{Find } \phi_5: \quad \phi_e = e\phi$$

$$\phi_5 = 5(2.02)$$

$$\phi_5 = 10.10 \text{ radians}$$

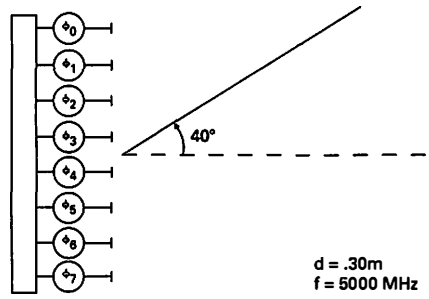


FIGURE 4-39 *Example 4-5*

Note that this result is greater than 2π radians. In practice, the phasers can shift the phase of the energy going to an element only by an amount between -2π and 2π radians. Expressed mathematically:

$$-2\pi \le \phi_i \le 2\pi$$

Thus, the phase shift applied to element five must be

$$\phi_5 = 10.10 \text{ radians} - 2\pi \text{ radians}$$
$$= 3.82 \text{ radians}$$

In other words, element five must lead the reference element (element zero) by an amount of 3.82 radians.

EXAMPLE 4-6: Hypothetical three-dimensional search radar. In this example, azimuth information is obtained by rotating the antenna mechanically in a continuous full-circle search pattern; see figure 4-40. Range information is obtained in the standard way by timing the pulse travel to and from the target. The phase shifters (0), (1), (2), (3) control the elevation position of the beam. It is desired to control the elevation of the beam in 0.0872 radian (5°) steps from +1.047 radians (60°) to −0.2617 radian (−15°) with respect to the antenna boresight axis, which is +0.2617 radian (15°) displaced from the horizontal. The system operational parameters are as follows:

Antenna rotational speed—10 rpm
Pulse repetition rate (PRR)—400 pps
No. of elevation beam positions—16
No. of pulse per beam position—2

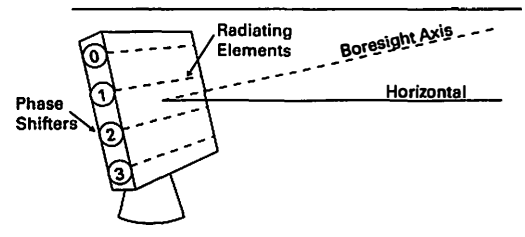


FIGURE 4-40 *Three-dimensional search radar employing phased-array techniques in elevation*

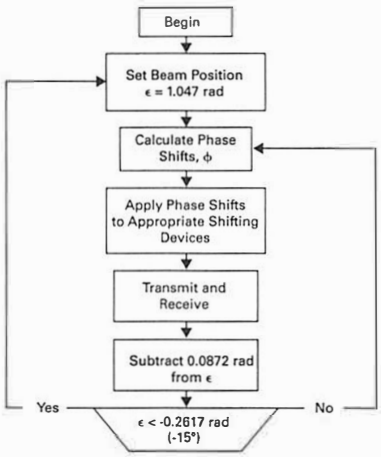


FIGURE 4-41 *Elevation beam positioning logic (for antenna whose unshifted beam center is 15° above the horizon)*

SOLUTION: For each beam position, the amount of phase shift is calculated for each radiating element. The resultant phase shifts are applied, and then two pulses are transmitted and received. The elevation scan logic is illustrated by the flow diagram in figure 4-41. The resultant scan pattern is illustrated in figure 4-42.

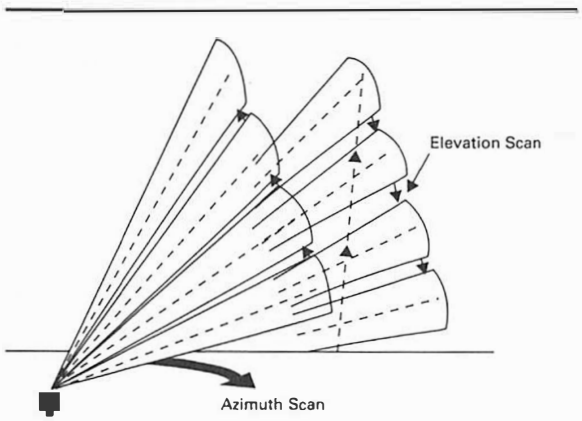


FIGURE 4-42 *Scan pattern for three-dimensional radar. For simplicity, all beam positions are not shown, and scanning angles have been exaggerated*

This example is, of course, hypothetical. Some operational three-dimensional radars, such as the Marine AN/TPS-59, function similarly; see figure 4-43. It is important to note that the concept of controlling beam position by varying the relative phase of radiating elements is common to frequency-scanned, phase-scanned, and time-delay-scanned arrays. The difference is in the methods employed to achieve the proper phase relationships between the radiating elements.

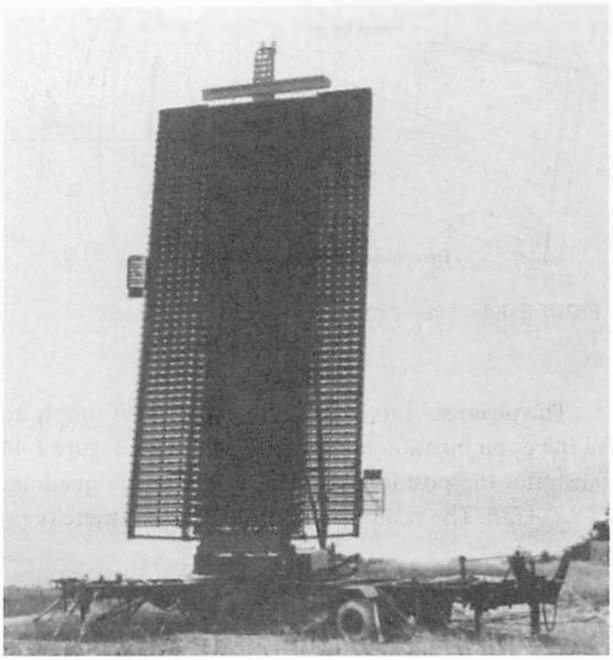


FIGURE 4-43 *The Marine Corps AN/TPS-59 three-dimensional radar. This radar is mechanically scanned in azimuth and uses phase shifters to scan in elevation.*

EXAMPLE 4-7: Full three-dimensional phased array-radar. The logical extension of the simple system of example 4-6 is the realization of a fully operational three-dimensional phased-array radar system to compute direct radar beams in elevation and bearing. Such a system is used in the Aegis system and in the Patriot missile acquisition and guidance radar.

The task now is to position the beam in both elevation and azimuth by electronic means. As in the two-dimensional case, the phase shift for each element in the array must be computed and applied prior to transmitting a pulse of radar energy.

SOLUTION: The array is made up of many independent elements. A unique element can be designated by the subscripts (a,e) for the ath column and eth row. The equations governing the positioning of the beam are presented below.

Azimuth to the Right or Left of the Boresight Axis (Azimuth scan)

$$\phi_a = a\frac{2\pi}{\lambda} d_a \sin\alpha \qquad (4\text{-}35)$$

Elevation Above or Below the Boresight Axis (Elevation scan)

$$\phi_e = e\frac{2\pi}{\lambda} d_e \sin\varepsilon \qquad (4\text{-}36)$$

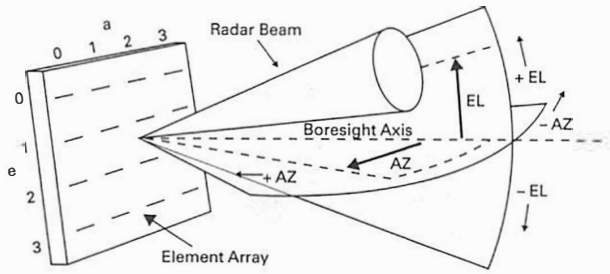


FIGURE 4-44 *Three-dimensional phased-array radar*

The phase shift for each unique element is simply an additive combination of the above equations. Figure 4-44 illustrates the positioning of the beam in the quadrant (+*AZ*, +*EL*). The combination of equations is therefore:

$$\phi_{a,e} = \phi_a + \phi_e$$

$$\phi_{a,e} = a\frac{2\pi}{\lambda} d_a \sin\alpha + e\frac{2\pi}{\lambda} d_e \sin\varepsilon \qquad (4\text{-}37)$$

$$\phi_{a,e} = \frac{2\pi}{\lambda}[ad_a \sin\alpha + ed_e \sin\varepsilon]$$

where:

$\phi_{a,e}$ = phase angle for elements *a*,*e* (radians)
λ = wavelength of the carrier (meters)
d_a, d_e = spacing between elements *a* and *e*
α = azimuth angle
ε = elevation angle

The example has been patterned after radars that use phase scanning in both directions (elevation and azimuth). Other systems have been designed using a combination of array scanning systems. For example the AN/SPS-33 radar, which is no longer in use, was frequency scanned in elevation and phase scanned in azimuth.

Active Electronically Scanned Array (AESA)

In our earlier discussion of radars, the antenna functioned solely as a conduit for passing the transmitted signal. Although the shape and size of an antenna might focus and thereby amplify or direct the signal, it did not alter the signal. Even with many current electronically scanned radars like the AN/SPY-1 phased-array radar, the source of the signal is still the transmitter. Although numerous phase shifters or phasers alter the phase of the signal, the antenna array still only passes the signal. This is called a passive antenna. Some more modern systems use what is called an active array.

An active antenna array actually generates the signal at the antenna. There is an array of transmit/receive (TR) modules on the face of the antenna in a similar arrangement to the multiple feed horns found in a conventional phased array radar; see figure 4-45. The modules are governed by the beam steering control unit (BSC) that directs the phase shift required to steer the beam.

Even though the power output of any one element is small, the additive effect of hundreds or thousands of these elements is significant. This configuration has several advantages over conventional mechanically scanned

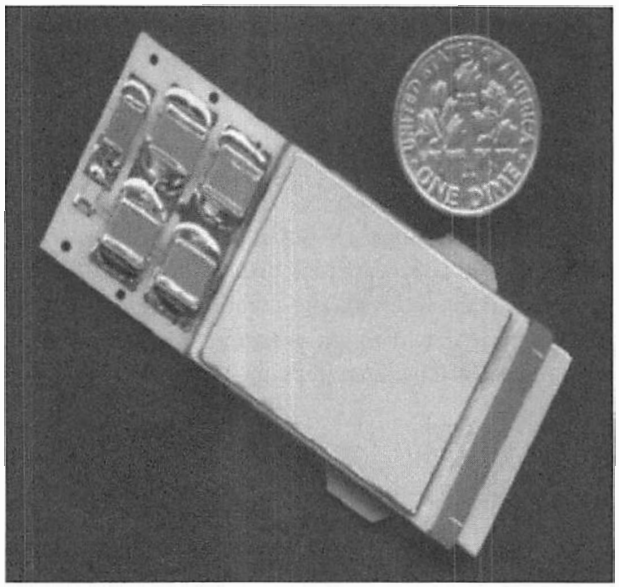


FIGURE 4-45 *Photo of AESA Transmit/Receive Module courtesy of Northrop Grumman Electronic Systems*

TABLE 4-2 *Summary of advantages of active electronically scanned array*

Reduced Weight	Elimination of wave guides and moving parts
Improved Reliability	Elimination/reduction of single-point failures
Graceful Degradation	Loss of one or several elements has minimal effect
More Frequency Agile (Bandwidth)	Elimination of wave guide and physical limitations
Higher Efficiency	No losses between transmitter and antenna
Higher Effective Power	Greater power compared to a similarly sized mechanically scanned antenna
Fast Beam Switch (Interleaved Modes)	Near instantaneous switching between air-air and air-ground or other modes
RCS Reduction	Antenna can be installed on a tilt and still function in order to avoid highlighting as compared to a vertically oriented antenna

antennas and passive electronically scanned arrays, as summarized in table 4-2.

Other Considerations

This has been a highly simplified treatment of electronic scanning and the phased-array radar. In addition to positioning the main lobe of energy, other considerations are:

1. Suppression of sidelobe interference.
2. Array element excitation amplitudes and geometry to achieve various beam radiation patterns.
3. Combining the phased array with a tracking system.
4. Modulating the radiated energy to transmit information—for example, in missile-guidance commands, and so on.

In practice, there are limits to the useful angular displacement of an electronically scanned radar beam. One limit is caused by the element pattern. The antenna pattern of an array is the product of the array pattern and the element pattern. In the simple examples given in this section, we have assumed that the element pattern was omnidirectional. A practical array element pattern is not omnidirectional, however, so the elements limit the scan angle.

Another limit is caused by the element spacing. A large scan angle requires close element spacing. If the scan angle exceeds that which can be accommodated by the element spacing, grating lobes will be formed in the other direction.

Effects of Changing Frequency

Millimeter Wave Radar

This topic is presented because of the increased interest in these systems in recent years and also as a demonstration of how one parameter (frequency) can affect the capability of a radar system.

Somewhat arbitrarily defined, millimeter waves range from 1 millimeter to 10 millimeter wavelengths, or from 300 GHz to 30 GHz, respectively. Examples of applications in which these high-frequency radars can be employed are in accurate image identification and resolution, in remote measurement, and in vehicle collision avoidance systems.

The extremely short wavelengths of millimeter radar cause certain fundamental operational characteristics to be more extreme. For example, practical millimeter radars have narrower antenna beamwidths than do microwave radars. This should allow better angle resolution with relatively small equipment. Small beamwidths imply not only high accuracy, but also reduced vulnerability to jamming (the enemy must detect the beam) and the ability to track at low altitudes without multipath interference from earth reflections.

Another way that radar performance is affected in this frequency domain is in the Doppler shift from moving targets. Equation 4-3 shows that as the wavelength decreases, the Doppler frequency shift will increase. This makes it easier to detect targets with small relative speeds.

However, one major limitation of millimeter wave radar is that high frequencies are much more susceptible to atmospheric absorption and propagate a fraction of the distance compared to lower frequencies. In fact, the absorption is so severe that to obtain reasonable propagation ranges (20 kilometers), the operating frequency must be selected from bands centered at approximately 35, 95, 140, or 220 GHz. These regions are called propagation "windows," which will be discussed further in chapter 7.

Within these windows, propagation through dry contaminants such as dust and smoke is good as compared to that in infrared systems. However, this absorption "limitation" may also be used to an advantage. For example, if short-range operation is required to avoid detection, then a frequency outside a transmission window may be suitable.

This discussion of millimeter radar is not complete, but we hope it has provided the reader with some appreciation of how a single parameter can change the operational capability of a radar.

Over-the-Horizon (OTH) Radar

At the other end of the radar frequency spectrum from millimeter waves are lower frequency radars in the HF region that have wavelengths on the order of 100 meters. Though the band between 1.8 and 40 MHz is subject to high noise levels and interference from long-range communications, it can provide detection ranges up to several thousand kilometers with the sky wave propagation path or several hundred kilometers when the ground wave path is used.

Sky wave radar Use of refraction layers in the ionosphere can extend radar range to 4,000 kilometers on a single hop; however, the long wavelengths required dictate the use of antennas 300 meters or more in length. Obviously this limits OTH radar to fixed land-based

sites if the entire HF portion of the spectrum is to be employed.

The actual range is determined by the frequency and elevation angle of the transmitted beam. The ionizing layers in the atmosphere change with time of day, season, and solar activity, which can result in multiple propagation paths or no refraction at all, degrading or eliminating employment of the radar in this mode. Elevation beamwidths must be very narrow in order to avoid multiple path propagation; however, this requirement tends to increase the already-large vertical dimension of the antenna.

Precise selection of frequency can allow the operator to select the specific refracting layer desired. Because of the large angle of incidence, considerable backscatter results when the beam returns to earth. This clutter is stronger than the returns from all aircraft and most ship targets. For this reason, the OTH radar must employ MTI, pulse-Doppler, CW, or FMCW methods to allow the target to be resolved in velocity and to reject nonmoving land or sea clutter.

Detection of ships requires more sophisticated processing techniques and longer observation times because their velocities closely approximate that of sea return or weather. Multiple hop propagation will result in increased clutter and degraded operation, but the range is increased by 100% with each hop. Use of a single frequency under optimum conditions could illuminate a range interval 1,000 kilometers deep starting at 1,000 to 4,000 kilometers. Using multiple frequency transmissions can increase the depth of this range interval, or if conditions are poor, it may require as many as six frequencies in one transmission just to cover the 1,000 kilometers interval.

Ground-wave radar With slight modification, the radar just described can be employed at frequencies that propagate as ground waves. The beam is transmitted at zero elevation and has a detection range against low-flying aircraft of 200 to 400 kilometers. Again, the long wavelength employed restricts its use to shore facilities, but it is able to maintain surveillance down to the surface of the earth, assuming land and sea clutter rejection features similar to those employed in sky wave radars are used. In actuality, detection is easier than with sky wave radars, because the angle of impingement with the earth's surface is smaller and backscatter is less.

Summary

Although radar is by no means the ultimate solution to the target-sensing and detection problem, it has become an integral factor in the development of advanced weapon systems. Radar is extensively used not only for target detection, but also for target tracking and weapon guidance. It is not limited in suitability to any single environment and is equally effective on land, sea, and airborne weapon systems.

There are three basic types of radar systems in current use: pulse, continuous-wave, and pulse-Doppler/MTI. Pulse systems are of value because of their ability to measure target range, whereas the value of continuous-wave systems lies in velocity discrimination. Pulse-Doppler systems are essentially pulse systems that are modified to be sensitive to the Doppler shift of the return pulse. Another type of system that measures range uses a frequency modulation technique and is widely employed as a radar altimeter in modern aircraft.

Continuous-wave radar systems have the principal advantage of being able to selectively output for display targets that meet certain velocity characteristics. This selection is possible because the CW radar components allow for the comparison of transmitted and return signals and because an amplifier can be tuned to accept only the Doppler frequencies of particular interest.

Pulse-Doppler and MTI radars offer the features of range determination and velocity discrimination. Both are sensitive to the Doppler shift within returned pulses, and both make use of the fact that the amplitude of returned pulses from targets with no relative velocity will not vary with time. Thus, a fluctuating output from a pulse-phase detector would indicate a moving target.

Pulse-Doppler radars use a series of frequency gates to process the phase detector output, and older MTI radars use a delay-line canceler. MTI radars are normally search radars with a low PRF, whereas the high-PRF pulse-Doppler radars are more suited for fire control. Usually pulse-Doppler radars have duty cycles above 10%, but MTI radars have duty cycles considerably less than 10%.

Two problems inherent to these systems are blind speeds and range ambiguity. Blind speeds can be eliminated by making the PRF high enough so that the slowest blind speed is higher than a target could exhibit. Higher PRFs mean lower unambiguous ranges. This problem can be counteracted by frequency agility in transmitted pulses, enabling the "tagging" of a particular pulse to be used for range measurement. Finally, a staggered PRF can be used as a solution to both the blind speed and ambiguous range problems.

Electronic scanning and its application to a phased-array system have been presented, focusing upon the concept of beam positioning. The basic concepts of radar

beam steering as a result of phase differences among multiple radiating elements were presented along with equations that determine the direction of a resultant beam. It should be noted that even with the increased complexity of the system, the output of the phased-array radar is the same as any other basic sensor, that is, target position. The great advantage of an electronically scanned system is that a single radar can perform multiple functions previously relegated to several separate radar systems. Phased-array technology is also being applied to nonradar systems such as IFF, communications, electronic combat, and SONAR.

---

[i] Hall, 105
[ii] Hall, 106
[iii] Hall, 108
[iv] Hall, 108
[v] Sullivan, 315
[vi] Hall, 114
[vii] Hall, 112
[viii] Hall, 113
[ix] Hall, 115
[x] Hall, 116
[xi] Sullivan, 23
[xii] Hall, 118
[xiii] Hall, 119
[xiv] Hall, 120
[xv] Hall, 121

5 Tracking Systems

Introduction

The purpose of a tracking system is to determine the location or direction of a target on a near-continuous basis. An ideal tracking system maintains contact and constantly updates the target's bearing, range, and elevation. Output of the tracking system is then sent to a fire control system, which stores the information, derives the target's motion, and then derives its future position. Tracking systems not only provide an automatic target-following feature but also determine the target's position with sufficient accuracy for weapon delivery.[i]

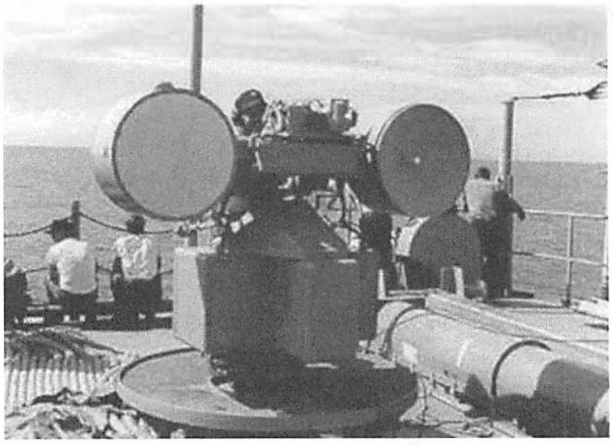


FIGURE 5-1 *Basic Point Defense Missile System; fire control system is human controlled.*

In the past, human judgment alone was relied on. Gunners on ships used what is commonly described as "Kentucky Windage" to calculate a firing solution. This progressed to actual gun sights that could be set or adjusted to precomputed lead angles for appropriate targets. These lead angles were condition specific and the operator was still required to track the target.

As the Navy moved to what today is viewed as crude analog fire control computers, great strides in accuracy were made, but there was still a requirement for much human intervention. Even moving into the missile age, manually guided systems such as the Basic Point Defense Missile System used a human operator to physically aim the missile illuminator; see figure 5-1.

Clearly this was not a desirable solution. Eventually, systems evolved into a mechanical assist or servomechanism where guns, missile launchers, and radar antennas are electromechanically controlled with inputs from a fire control computer.

Servo Systems

While the electronically scanned array has become more common, there are still many systems that rely on mechanically trained antennas to maintain the line-of-sight (LOS) between the sensor and target. Some examples are surface-to-air missile illuminators, the Close-In Weapon System, and the Automatic Carrier Landing System.

The primary issue with a mechanically trained system is error correction. If the tracking element at all times were pointed directly along the line-of-sight, tracking would be perfect and no error would exist. In reality, error is always present, and therefore a second line, called the tracking line, is defined as the line that forms the axis of symmetry of the radiated energy, commonly called the antenna boresight axis; see figure 5-2.

When error exists, the line-of-sight and the tracking line are not coincident; usually the tracking line lags the line-of-sight by some small amount, due to the general inability of mechanical tracking systems to anticipate target movements. This inability defines the central problem of any tracking system: How can the error between the line-of-sight and the tracking line be reduced to an acceptable minimum average value?

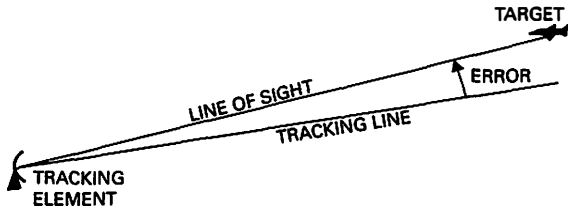


FIGURE 5-2 *Relationship between the line-of-sight and the tracking line*

The problem of minimizing error between the LOS and the tracking line is further complicated because weapon platforms are not generally stable. The combined effects of weapon platform roll, pitch, and yaw produce a disturbance in the tracking system that is not strictly related to target motion.

These rotational motions not only affect the tracking element's ability to hold the target, but also generate erroneous output data. Since the basis of relative target velocity computation is an accurate measurement of target position over time, the uncompensated rotational motions of the weapon platform result in erroneous target velocities being generated. As will be seen in a later chapter, target velocity data is a major input for the computation of weapon launch lead angle.

One of the most basic tracking system designs is the servo tracking system. Servo tracking systems require separate components for range and angle tracking of a specific target. These functions were performed manually in early radars by operators watching target returns, and then positioning handwheels to maintain markers over the desired target's return "blip" on a scope for azimuth, elevation, and range.

The handwheels positioned potentiometers, which adjusted voltages analogous to target azimuth, elevation, or range, which were supplied to fire control computers and lead-computing gun sights. As target speeds and operator workloads increased, automatic tracking circuitry was developed that maintained tracking efficiency over sustained periods and made possible later developments of unmanned homing-type weapons.

A simple servo tracking mechanism is shown in figure 5-3. Here, the radar antenna is initially aimed at a target, after which the antenna automatically follows the target's motion. Furthermore, the system provides continuous position information to an operator and possibly to a fire control system. As shown, the antenna is physically rotated by a motor that gives a continuous negative position feedback signal to a controller. The subsystem that controls the direction of the antenna is called a servomechanism; see figure 5-4.[ii]

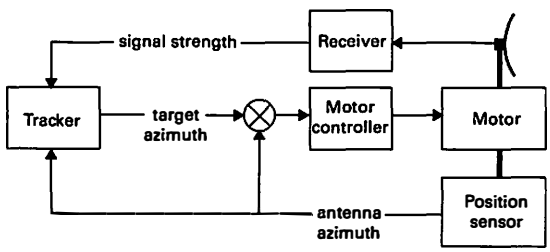


FIGURE 5-3 *Servo tracking system*

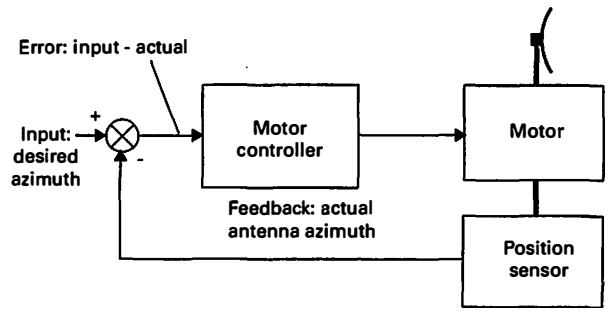


FIGURE 5-4 *Servomechanism*

The input signal to the servomechanism is the desired azimuth of the antenna. The error signal drives the motor to reposition the antenna until the position feedback indicates that the antenna is at the desired azimuth, at which point the error signal is zero and the motor stops. This servomechanism can be combined with a tracker, which determines the azimuth to the target, becoming the new system input.[iii]

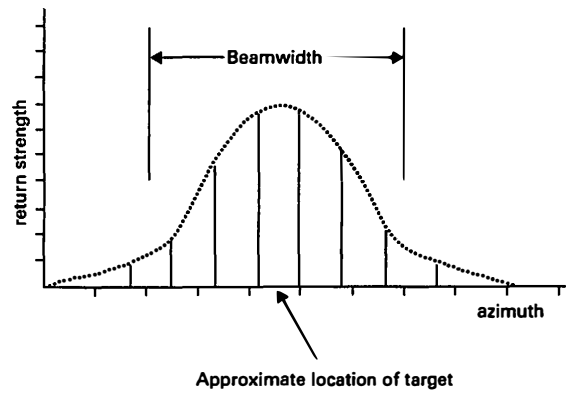


FIGURE 5-5 *Return strength as beam sweeps over target*

It must be mentioned that the diagrams are concentrating on azimuth. There are equivalent components in elevation as well. The "brains" of this system is the tracker, which takes the return signal and position information and determines the location of the target. There are several ways to accomplish this, with varying degrees of complexity and accuracy. Consider the following data in figure 5-5, which might be obtained as a circular sweep radar pans across a target.[iv]

Individual returns are shown as vertical lines. As the beam sweeps across the target, there are eight generated returns whose strength increases until the target is centered in the beam and then falls off. Target location can be determined from the point of maximum return strength.

However, the region of maximum return is broad and therefore the exact location of the target cannot be determined with high accuracy. Certainly, one could locate the target within some fraction of the beamwidth, but even with this system, the tracking accuracy would not be much better than about one-quarter of the beamwidth. Because this would be insufficient for weapon delivery, there is need for more accuracy.[v]

The concepts discussed here are applicable to a large group of diverse systems including shipboard fire control monotrack systems, homing missiles, acoustic homing torpedoes, and aviation fire control tracking systems. In general, all tracking servo systems function conceptually in the same manner, accomplishing the following functions:

1. Sensing position error magnitude and direction
2. Providing position feedback
3. Providing data smoothing/stabilization
4. Providing velocity feedback to aid in achieving a smooth track
5. Providing a power driving device

This discussion will be developed from the viewpoint of how the servo system illustrated in figure 5-6 satisfies each of functions 1 through 5 above.

Stabilization

It should be understood that no single method of compensation for rotational movement exists. Nevertheless, all systems incorporate gyroscopic devices in one way or another. With reference to stabilization, tracking systems can be grouped into three major classes:

1. *Unstabilized*—The tracking subsystem operates in a totally unstable environment, and therefore its output contains rotational motion components. Gyroscopic compensation is external to the tracking subsystem.
2. *Partially stabilized*—The tracking subsystem is stabilized along one axis for example cross level, and the output contains rotational disturbances associated only with the uncompensated axis. The remainder of compensation is external to the tracking subsystem.
3. *Fully stabilized*—The tracking subsystem is gimbal-mounted and remains totally unaffected by rotational movement. Its output is completely free of rotational disturbances, and therefore no further compensation need be done prior to computation.

The reference gyroscope (gyro) can be dedicated to a specific weapon system, or all weapons and sensors could be served by a single gyro reference. The usefulness of a gyro as a stable reference is due to its tendency to remain in a fixed plane in space if no force is applied to it, and its tendency to turn at right angles to the direction of an externally applied force. Chapter 13 provides greater details of gyro functions and properties.

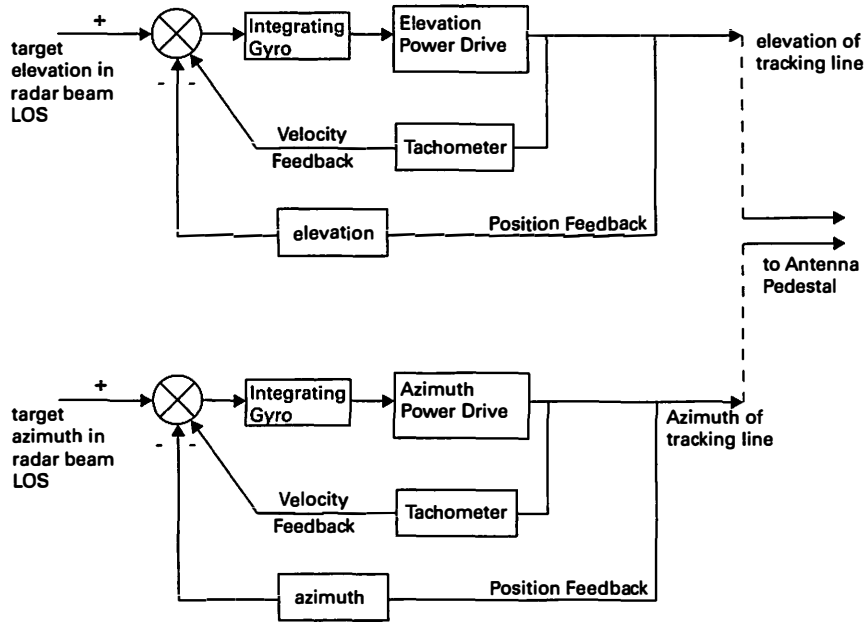


FIGURE 5-6 *Block diagram of a typical automatic-tracking system in azimuth and elevation, employing a radar sensor system*

Position Error Magnitude and Direction

This function is performed by the radar itself in conjunction with circuitry designed specially for the type of radar employed. From the study of radars, recall that generally the output of a sensor is target position data, and these data are points available only when the sensor is pointing at the target. Since target position data should be available to the weapon control system at all times, some means of detecting target motion is required if the sensor is to follow the target.

Recall also that target information enters the sensor from a beam of energy, with the energy concentrated along the axis of the beam. For example, using a radar sensor, target information is available when the target is within the radar's beam of energy and is strongest when the target is on the beam axis. As the target moves to the periphery of the beam, the return signal level begins to diminish. The amount of signal reduction can therefore be used as a measure of how far away the target is from the beam axis and ultimately can be employed to position the radar antenna back on the target automatically.

However, the return signal level will diminish by the same amount regardless of which direction the target takes with respect to the beam axis. For this reason, a means of sensing direction is required in addition to that of sensing magnitude. If the target is not exactly in the center of the radar beam, then an error will exist between the direction that the radar antenna is pointing and the true line-of-sight to the target.

Azimuth Tracking

There are many different methods of tracking a target on the basis of its position in the radar beam. The following discussion concentrates on some of the most commonly used concepts.

Sequential lobing One of the first tracking methods developed was lobe switching or sequential lobing, which is still used in some countries. In this method a beam is positioned in one direction, and in subsequent pulses the beam is orientated slightly offset from the previous direction. Direction in one axis can be determined by a double-lobe system very rapidly. If the angular error in the orthogonal coordinate is desired simultaneously, the beam could be stepped in a minimum of three (usually four) positions for signal comparisons in a process called sequential lobing; see figure 5-7.

The stepping process is limited by the PRF of the system because the beam position must be stopped long enough for at least one transmission and received echo

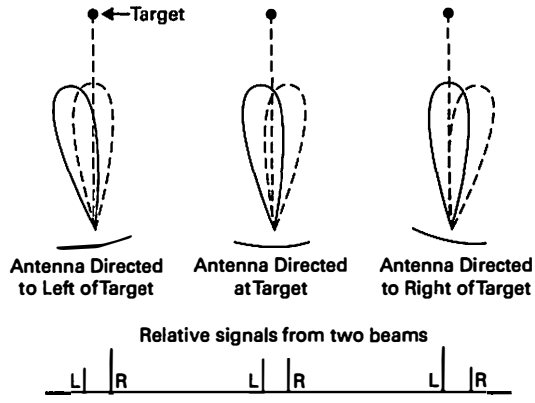


FIGURE 5-7 *Sequential lobing in azimuth*

cycle, and probably more prior to switching to the next position. Four separate antennas or four separate feed horns are required in addition to very complex wave guide plumbing and switching techniques. Because three horns or antennas are turned off at one time, considerable scanning losses occur, and the data rate is extremely low unless sophisticated electronic lobing systems are used.

Conical scan The natural extension of the sequential lobing process in four positions is to rotate the beam continuously in a circular fashion. Error between the tracking axis and the line-of-sight can then be sensed in magnitude and direction by causing the radar beam to be rotated rapidly about the axis of the antenna, thus producing a narrow cone of energy in space; see figure 5-8.

This is generally accomplished mechanically by nutating (rotating) the feed horn in a small circle around the focal point of a fixed paraboloid; thus, the antenna beam will lie along the axis of the reflector. If the feed point is

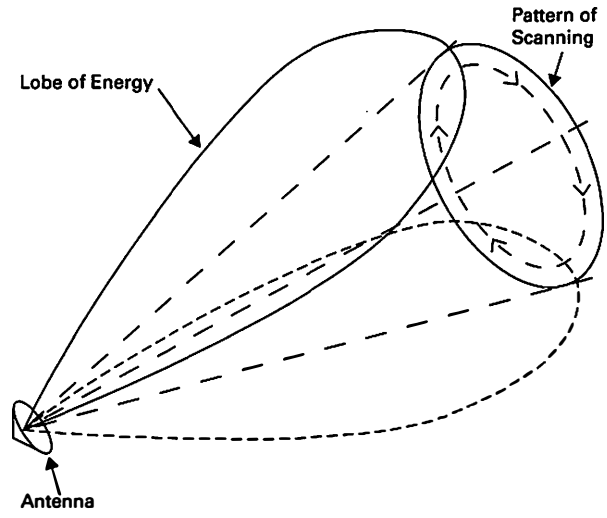


FIGURE 5-8 *Conical scanning*

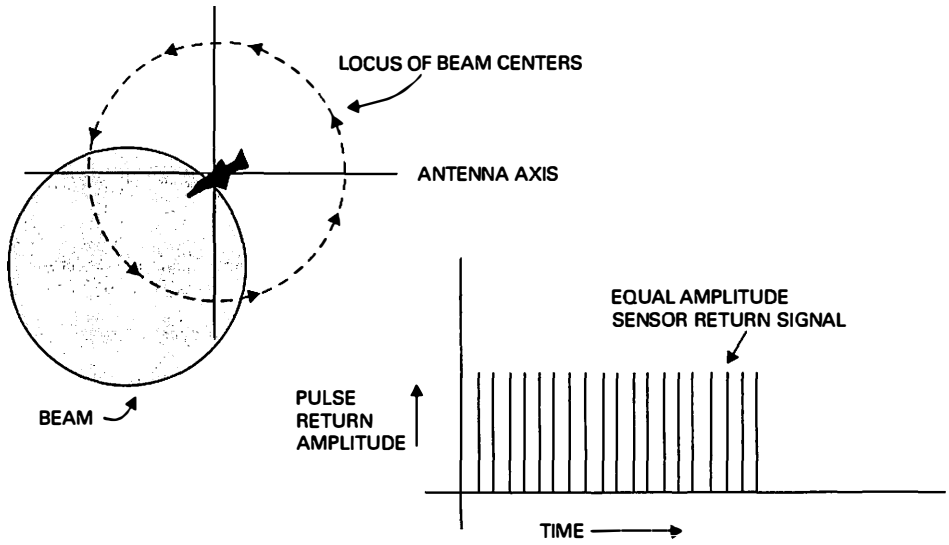


FIGURE 5-9 *Relationship of target position and return pulse amplitude when the target is on the tracking line*

moved transversely slightly away from the focus, the beam will be at an angle with the axis. If the feed point is oscillated back and forth, the beam will swing from side to side. If the feed point is nutated in a circle about the axis, a conical scan will result.

Now examine figure 5-9. In this case, the antenna axis is pointed directly at the target, so no matter where the transmitted lobe is in the nutation cycle, the echo received by the radar is always of the same amplitude. Although maximum possible energy return is never received, the resulting echoes are all equally strong, and target tracking is very accurate.

The rate of beam rotation must be such that several pulses of radiated energy are exchanged with the target during each period of beam rotation. In actual practice, the ratio PRF/f_s is approximately 40 to 1, where f_s is the scan frequency and PRF is the pulse repetition frequency of the radar.

When the LOS and the tracking line are coincident—that is, the target is exactly on the tracking line—successive target echoes received by the system will be of equal amplitude since the beam center is at all times equidistant from the target during each scan period. However, if the target is not exactly on the antenna axis, an error exists between the line-of-sight and the tracking line. This error is immediately detectable because as the beam rotates, the return pulses are amplitude modulated due to the varying proximity of the target with the beam center. The frequency of modulation is the scan frequency of the system f_s.

This concept of amplitude modulating the sensor input signal is not limited in application to active radar systems. Any system whose function is to track an energy source can be made to do so by employing the conical scan concept—that is, laser, infrared, and acoustic sensors.

Using the error signal generated from the amplitude modulated input signal shown in figure 5-10, the radar will be repositioned to cause the returning input signal to again equalize as in figure 5-9.

For a system using conical scan techniques, actual movement of the beam through space is easily detected by the target. Since the error signal is computed directly from the varying amplitude of the return energy, it is possible for the target to use a transponder, which samples the incoming energy and then rebroadcasts it.

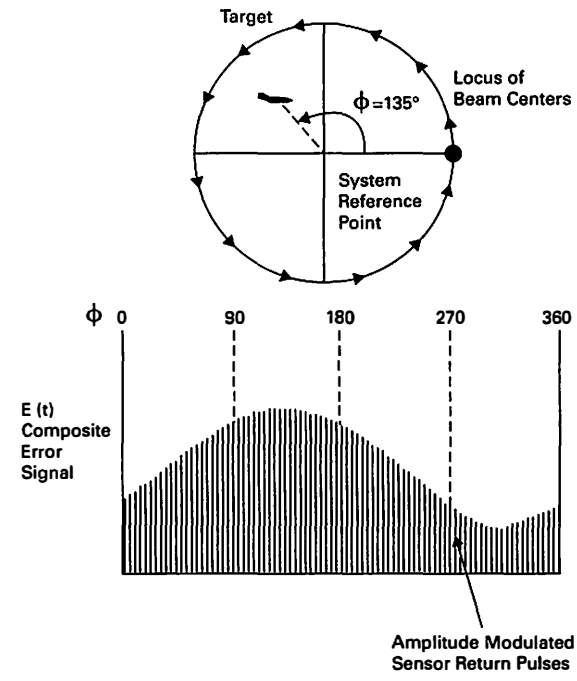


FIGURE 5-10 *Relationship of target position and return pulse amplitude when the target is displaced from the tracking line*

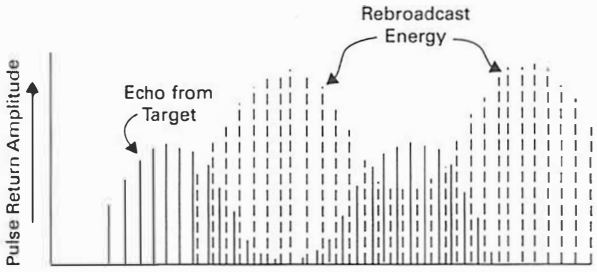


FIGURE 5-11 *Returning "skin" echo with jammer signal superimposed*

When the target senses maximum incoming energy, the transponder rebroadcasts an amount equal to or slightly greater than the energy sensed. As the energy falls off in amplitude due to the major axis of the beam moving away, the transponder will broadcast a signal with proportionally greater amplitude; see the dotted lines in figure 5-11.

When the incoming energy increases with the approach of the main beam axis, the transponder will broadcast a signal with less and less enhancement, until at maximum it is again equal to or slightly greater than the radar echoes. As a result, the radar's tracking system drives the radar away from the target. This technique is referred to as inverse conical scan and is a common means of deceiving conical scan tracking systems.

COSRO As a means of achieving a degree of immunity to deception from inverse conical scan, a system known as COSRO was developed. COSRO, an acronym for conical scan on receive only, employs a nonscanning transmit beam and a rotary (nutating) scanner in the wave guide between the duplexer and the receiver in order to extract elevation and azimuth angle error. A radar warning receiver in a target being tracked by a COSRO-type radar would detect a steady, nonscanning pulse radar beam that would provide no information to assist in determining deception techniques beyond the radar performance parameters listed in chapter 3.

During rest time between pulses, the duplexer switches the antenna to the COSRO scanner and receiver. The COSRO scanner has the effect of shifting the axis of greatest sensitivity to echo reception, similar to the slight change in the direction of the transmitted beam of a conical scan radar caused by nutation of the feed horn.

One can imagine a receive "beam," similar in shape to a transmitted beam, being nutated about the radar boresight axis. The COSRO scanner determines the distribution of signal strength in the received energy by use of an RF signal detector rotated by a motor tachometer. The location of the greatest amount of returned energy is determined, and signals are sent to elevation and azimuth power drives to reposition the antenna on the target's center. When the target is centered, the COSRO scanner will sense equal signal strength throughout the rotation of the RF signal detector.

This concept can also be employed with sequential lobing radars to produce a lobe on receiver only, or LORO, radar. While there are techniques to deceive these types of radars, much more sophistication, complexity, and expense are involved.

Monopulse Conical scanning and sequential lobing techniques determine the target's position relative to the radar's main beam axis. If the target is on axis, its amplitude for the return pulse is maximized. If the target is off axis, the return pulse's amplitude is less. The amount of decrease is dependent on how far off axis the target is.

Using this idea, a train of pulses is sent out and their subsequent return amplitudes are measured at the receiver. An error signal is extracted from the difference between pulses and is used to position the radar.

Serious tracking errors can be generated in such a system if the return of any one pulse is markedly increased or decreased in amplitude as compared to the next pulse in the train. Such a fluctuation could easily occur if there were a sudden change in the target's radar cross section (or any number of other reasons).

To avoid these errors, the U.S. Naval Research Laboratory (NRL) developed the monopulse radar in 1943; see figure 5-12. The monopulse technique was first applied to the Nike-Ajax missile system, which at the time was the continental U.S. air defense system. In 1958, this type of

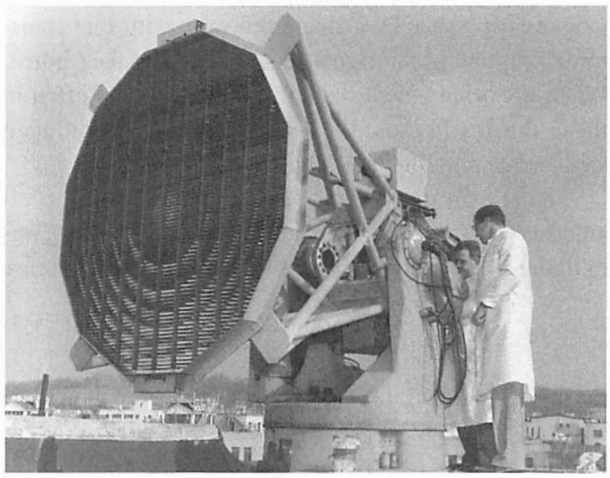


FIGURE 5-12 *An experimental monopulse radar system. To overcome the angular limitation of existing radars, the monopulse radar, developed by the NRL, provided a tenfold improvement in angular accuracy over that previously attainable in the training and pointing of fire and missile control radars.*

radar was also used to guide the launchings of the first U.S. space satellites at Cape Canaveral. Monopulse radar is now the basis for most modern tracking and missile servo-controlled tracking radars because of its high accuracy and relative immunity to electronic countermeasures that degrade other tracking methods.

A monopulse tracking system determines amplitude modulation using only one pulse with multiple processed beams instead of a train of pulses.

To accomplish this, the target must be located in two or more beams simultaneously and a comparison made between the return pulse in each beam. Since the tracking error signal is derived from a single pulse, this technique is known as monopulse. The comparison of the return echo in each beam can be done in either amplitude or phase. The following discussion concentrates on the amplitude comparison.

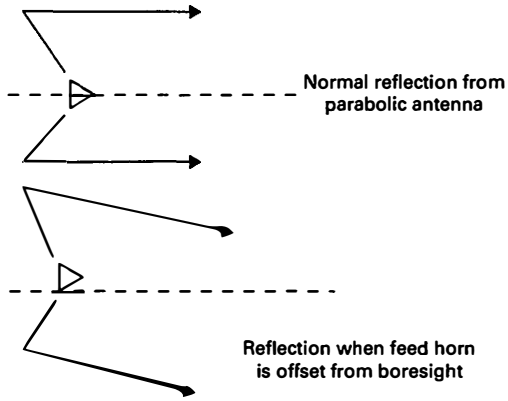


FIGURE 5-13 *How offset feed changes beam direction*

In a two-beam system, each beam is generated simultaneously in space by a single reflector, with one transmitter supplying the signal to two adjacent feed horns. When the beam is reflected off the reflector, the offset in the feed axis will cause the beam to be reflected off at an angle relative to the boresight; see figure 5-13.

Figure 5-14 shows these beams slightly overlapped. If an observer were to actually move across the beam paths in figure 5-14 with a power meter, the meter would sense

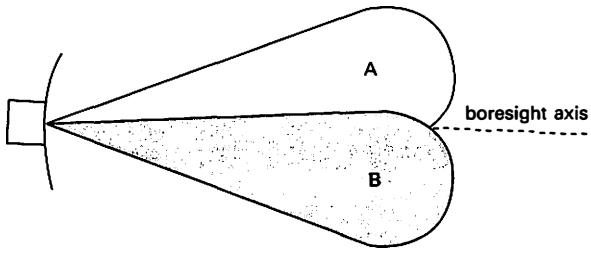


FIGURE 5-14 *Dual beams*

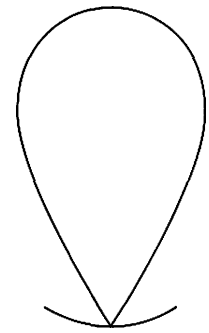


FIGURE 5-15 *Sum pattern of two beams*

the algebraic sum of the electromagnetic energy produced by the two beams. If the power meter were moved about to various positions in front of the reflector and the measured power levels graphed, a plot as depicted in figure 5-15 would be produced. Notice that the radiation pattern appears as one lobe even though it is really the sum of two separate lobes.

Monopulse radar is designed so that each beam can be identified separately by employing different polarization or some other tagging process. For this reason, the beams can be examined together, as depicted in figure 5-15, or separately. By summing the beams, the monopulse radar will receive a large return from a target centered between the two beams just as if it were tracking with the center of a strong single beam. This aspect is employed by the monopulse radar for target detection and range determination.

For angle tracking, the radar receiver will invert the value of the energy from one beam, thus changing the sign—see figure 5-16—and sum that negative value with the positive value from the other beam. If the target is on

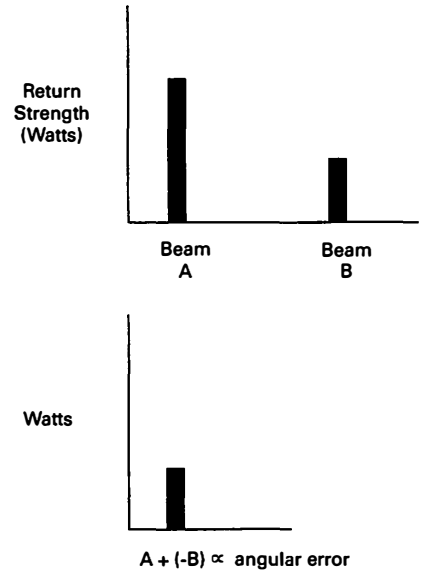


FIGURE 5-16 *Dual-beam angular output determination*

the boresight axis of the reflector, there will be equal reflected energy from the two beams. Thus, their algebraic sum would be zero and the tracker would be considered "locked on." If the target were to the left of the boresight axis, the sum would be positive, and if it were to the right, the sum would be negative.

The amount and direction of angular error in one axis is given by the magnitude of the error signal. The farther from centerline the target is, the greater the magnitude of the signal. This can easily be used as an error signal to drive a servomotor to keep the antenna pointed at the target, thereby reducing the error to a very small value.

Figure 5-17 depicts a simplified microwave comparator circuit of a monopulse system in both azimuth and elevation. The microwave comparator circuit is composed of wave guide and hybrid junctions called magic tees.

The hybrid junctions have two inputs and two outputs. One output is the sum of the inputs, and one is the difference of the inputs. Four hybrids can be connected as shown to provide azimuth and elevation error. In this case, a cluster of four feed horns producing four beams, indicated by the letters *A*, *B*, *C*, and *D*, are used.

To determine azimuth, the sum of the *C* and *D* horn outputs are subtracted from that of the *A* and *B* horns. If the result is positive, then the target is to the right of the boresight axis, assuming the horns at the top of the diagram are facing the reader. If the result is negative, then the target is to the left of the boresight.

Elevation error is determined by subtracting the sum of the outputs of the *B* and *D* horns from the sum of the outputs of the *A* and *C* horns, a positive value indicating that the target is above the boresight and a negative value indicating that the target is below the boresight. The tracking circuit and power drives attempt to equalize the amplitude of the target echo in all four horns that resulted when the target was on the boresight axis.

The signal-to-noise ratio for a monopulse tracker would generally be higher than that of a conical scan system, since the target is tracked on the axis of summation of the beams rather than on the edge. The target would not be able to sense whether it was being tracked by a monopulse or a COSRO radar, thus making deception jamming techniques complicated.

It is very difficult to deceive a monopulse radar in angle, but it is as easy (or as hard) to deceive it in range as with any other radar. The disadvantages of increased complexity and cost of monopulse over that of other techniques are being reduced by advances in technology, therefore most of today's servo tracking and homing systems incorporate some variation of this system.

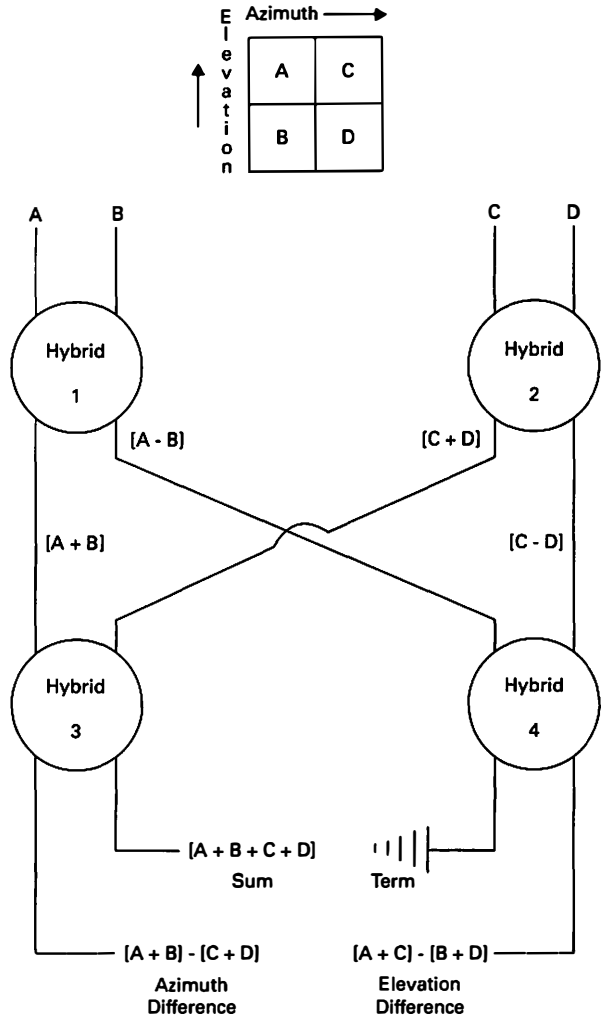


FIGURE 5-17 *Monopulse radar signal comparator circuit*

Position Feedback

When an error signal is developed and the system is caused to move in response to it, the system must know when it has reduced the error near zero. The position feedback is an intrinsic part of the error signal generation from the radar. As the system moves in response to the original error, the result is to position the tracking line in coincidence with the LOS. This action by the system reduces the position error signal near zero, thus providing an indication that the system has responded correctly to the error initially measured by the radar.

This process is essentially instantaneous and continuous throughout the entire time that the sensor is providing target information to the system. The equilibrium state of the tracking system then is a null composite error signal, resulting in zero output to the drive system. It must be understood that the true equilibrium state is *never* achieved

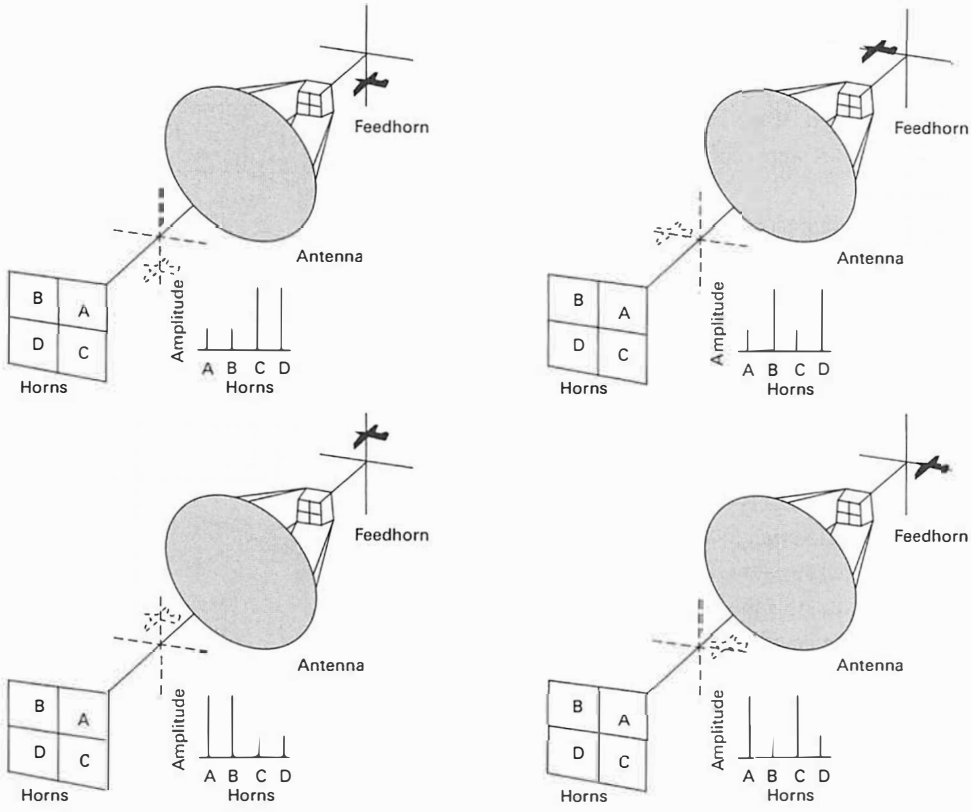


FIGURE 5-18
Amplitude changes of received energy with target position

while the system is actually tracking, but that if operating properly, the system will always tend toward this state.

Power Driving Devices

Naval weapon systems generally employ two broad categories of motive devices: electric motors and electrically controlled hydraulic motors. These devices are used to actually move the driven element of the system. Which of the two types of motors to use depends upon the physical size of the load to be driven. Smaller loads, such as missile-steering control surfaces and small fire control directors, are driven by electric motors. Electrohydraulic motors are used to position heavy loads such as missile launchers and gun mounts.

Velocity Feedback

The primary purpose of velocity feedback is to aid in the prevention of dynamic overshoot. The velocity feedback signal is employed in a degenerative manner—that is, the feedback voltage is subtracted from the output drive signal of the gyro. This subtraction serves to limit the speed of the output motors at an equilibrium level. This equilibrium level is governed by the ratio of the feedback voltage to the input drive voltage.

Range Tracking

Automatic range tracking has been an objective since the early days of radar. The first range-tracking units, developed during World War II, were electromechanical devices such as the Mechan Range Unit. As target closing speeds increased, electronic analog, digital, and hybrid rangers replaced mechanical units to respond to the threat.

The purpose of the range unit is to follow the target in range and provide continuous distance information or slant range to the target. Timing pulses are sent to the receiver and display unit to provide a range gate and allow the angle tracking system to look at only one specific range interval (really, a time interval) where the desired target echo pulse will be received. The locations of the range gate and that of the target are examined, the amount and direction of error between the center of the gate and the target is determined, error voltages are generated, and the circuitry responds by moving the gate to center it on the target. This operation is carried out by a closed-loop feedback system similar to the angle tracking servo system discussed previously.

There are several methods of sensing range-tracking error; however, all implementations involve some method of incrementing the distance to the maximum range of the radar such that specific portions of range interval

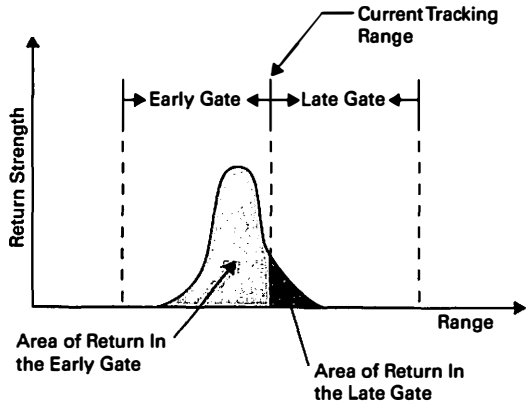


FIGURE 5-19 *Range gates*

can be examined. One type of range gating system is described below.

Range tracking is accomplished in a similar manner to dual-beam angle tracking. Once the range has been measured, the tracking system attempts to predict the range on the next pulse. This estimate becomes the reference to which the next measurement will be compared. Using two range windows, called the early and late range gates, makes the comparison; see figure 5-19.[vi]

The area of return in each gate is computed by integration. The difference between the area in the early and late gates is proportional to the error in the range estimate. If the two areas are equal, the return is centered directly on the range estimate, and there is no error. If the return has more area in the early gate, the range estimate is too great, and therefore the range error is positive. In the near vicinity of the range estimate, there will be a linear relationship between the range error and the difference in the areas; see figure 5-20.[vii]

As long as the range estimate is not too far off, the tracking error can be determined and the target range updated. Again, like dual-beam tracking, the range-tracking system can measure the target range with greater accuracy than the range resolution of the system, R_{res}, which is determined by the pulse width and possibly the pulse compression ratio.[viii]

Track-While-Scan Concepts

Various surface and airborne radars provide major sensor inputs to computerized command and control systems. These systems perform multiple functions, including surface, subsurface, and airspace surveillance; identification; threat evaluation; and weapon control. To operate in today's environment, command and control systems require near real-time (as it occurs) target position and velocity data in order to produce fire control quality data for weapon systems.

Until recent times, the interface between the raw data supplied by the sensor and the command and control system was the human operator; however, this limited the data rate and target handling capacity. Additional personnel were added as a partial solution to these problems, but even under ideal conditions of training, motivation, and rest, the human operator can update only a maximum of six tracks about once per 2 seconds, and that number for only a short time.

To supply fire control quality data, the classic solution has been to provide separate precision tracking sensors, such as the servo systems previously discussed. Although the systems have widespread use, they are limited to tracking at a single elevation and azimuth.

Data from these systems are provided to computers to solve the fire control problem and compute launcher and weapon orders. With the advent of high-speed digital computers, computational capacity exceeded the data-gathering capacity of the fire control sensors; however, target handling continued to be limited by the nonscanning tracking methods employed.

The solution to both problems was to interface a high-scan-rate (15-20 rpm) search radar with the digital

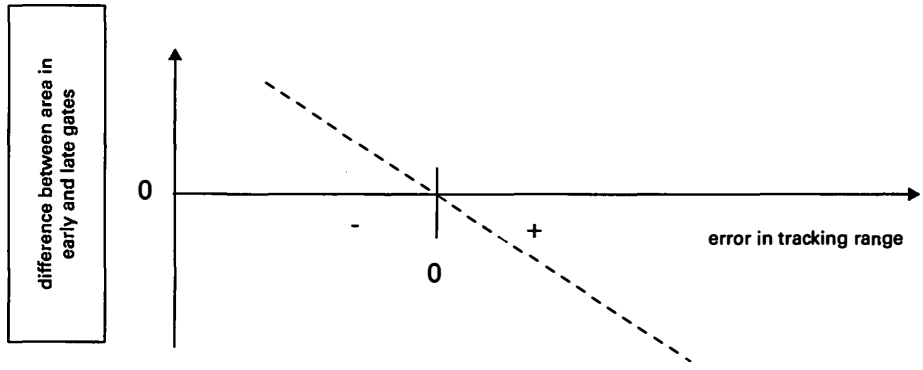


FIGURE 5-20
Range error

computer through a device capable of converting the pattern of echo returns detected by the receiver into usable binary information. In addition, using the digital computer to direct and compose radar beam patterns such as phase-array scanning created a flexible system capable of multimode, simultaneous operation in search, designation, track, and weapon control.

This ability to perform such functions as multitarget tracking, prediction of future target positions, and terrain avoidance for aircraft navigation has transformed previously fragmented sensor and control capability into a dynamic and powerful sensing, command and control, and weapon employment system. The combination of Automatic Detection and Tracking (ADT) functions within a search radar is sometimes referred to as track-while-scan (TWS). The central concept of TWS is to maintain target tracks in a computer with periodic information updates from a scanning radar.

Fundamentals of Automatic Detection and Tracking (ADT)

A functional tracking system must provide the basic target parameters of position and rates of motion. In the systems presented earlier in this text, both position data and velocity data were used to maintain the tracking antenna on the target at all times, thus limiting the system to traditional single-target tracking.

In a track-while-scan system, target position must be extracted and velocities calculated for many targets while the radar continues to scan. Obviously, in a system of this type, target data are not continuously available for each target. Since the antenna is continuing to scan, some means of storing and analyzing target data from one update to the next and beyond is necessary. A digital computer is employed to perform this function and thus replaces the tracking servo systems previously discussed.

A Tracking Algorithm

The following algorithm for solving the track-while-scan problem is based upon the assumption that the radar furnishes target position information once each scan. The scheme can be implemented with software only, or with a combination of special radar circuits and software. Existing systems also allow an operator to modify the system tracking parameters.

Any track-while-scan system must provide for each of the following functions:

1. Target detection
2. Generation of tracking "gates"
3. Target track correlation and association
4. Target track initiation and track file generation (if no correlation)
5. Track gate prediction, smoothing (filtering), and positioning
6. Display and future target position calculation

Target Detection

Target detection is accomplished by special circuitry in the receiver of radars designed as TWS systems or by signal-processing equipment in a separate cabinet. The functions of all ADT processing equipment are similar and will be addressed here in general terms.

It should be remembered that the ADT processor is not a computer, although it does have some memory capacity. Its primary function is that of converting data to a form usable by a separate digital computer that actually performs the smooth tracking, velocity, acceleration rate generation, and prediction functions of the TWS system.

The ADT processor's memory allocates space addressed to each three-dimensional spatial cell that makes up the search volume of the radar. The dimensions of these cells are often defined by the horizontal and vertical angular resolution capability of the radar as well as by a distance in range equal to the pulse compression range resolution.

After several scans, the ADT processor contains what amounts to a three-dimensional binary matrix representing the entire search volume of the radar. This matrix or array represents cells with echo returns with a "1" and those without returns with a "0"; see figure 5-21.

In some systems, individual beams in the radar are overlapped such that most targets will occupy more than one beam position at any radar range. When employed, this procedure, called beam splitting, allows angular resolution less than the beamwidth of the radar. In addition, this avoids gaps in coverage and allows examination of hit patterns among several beams as a means of rejecting clutter or false targets.

A radar may require N hits out of M beams at the same range interval, with all hits overlapping in elevation and azimuth before any further consideration is given to initiating a track. N and M may be operator selectable or preset during the design process.

Most systems employ further means to avoid generation of tracks on clutter or false returns. These can be grouped into those associated with conventional radar receiver antijamming circuitry and those exclusive to ADT. An effective TWS radar must retain conventional features such as constant-false-alarm rate (CFAR), MTI, and automatic gain-control circuitry. These help prevent the processing of obviously false returns, allowing the

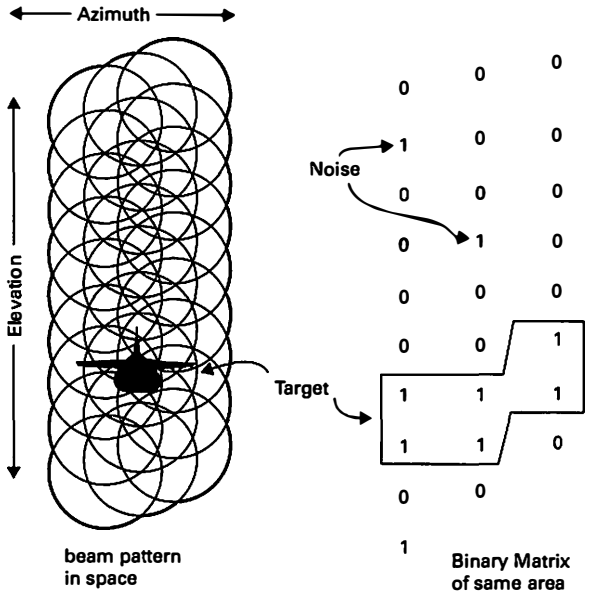


FIGURE 5-21 *The ADT processor converts echo-return patterns from successive positions of the radar beam into a three-dimensional binary matrix as shown. (*Note: *Only azimuth and elevations are depicted.)*

ADT system to make more sophisticated decisions between valid and invalid targets.

The ADT processor will generate a time history of all echo returns over several radar scans. Stationary targets remaining within a given resolution cell over a number of scans result in its identification as a clutter cell. The collected information on all clutter cells within the radar search volume is referred to as a clutter map. Returns remaining in clutter cells can be rejected as false targets. Moving targets will not remain in a given cell long enough to be eliminated.

Returns that very rapidly change in strength and centroid location, such as clouds and chaff in high winds, could appear nonstationary and would cause a track to be established. These returns are then tested for the known velocity characteristics of valid targets by velocity filters, thus reducing the number of false tracks that would otherwise be generated in this manner.

Target Acquisition, Tracking, and Turn Detection Gates

A "gate" can be defined as a small volume of space composed of many of the cells described previously, initially centered on the target, which will be monitored on each scan for the presence of target information. Gate dimension and position information is generated by a digital computer and sent to the ADT processor for application. Action by the clutter map is inhibited inside these gates to avoid generation of clutter cells by valid tracks.

When a target is initially detected, the algorithm receives only the position data for that initial, instantaneous target position. The acquisition gate is then generated in the following manner:

TABLE 5-1 *Example of gate dimensions from figure 5-22.* Note: *This is an example only; numbers will vary depending on the actual system.*

	Acquisition Gate	Tracking Gate
Range Gate	*R* 2,000 yds (1,829.15 meters)	*R* 120 yds (109.59 meters)
Bearing Gate	*B* 10° (0.1745 radian)	*B* 1.5° (0.0262 radian)
Elevation Gate	*E* 10° (0.1745 radian)	*E* 1.5° (0.0262 radian)

The acquisition gate is large in order to allow for target motion during one scan period of the radar. If the target is within the acquisition gate on the next scan, a smaller *tracking gate* is generated that is moved to the new expected target position on subsequent scans.

Although figure 5-22 shows only a very small tracking gate, in actual practice intermediate-gate size may be

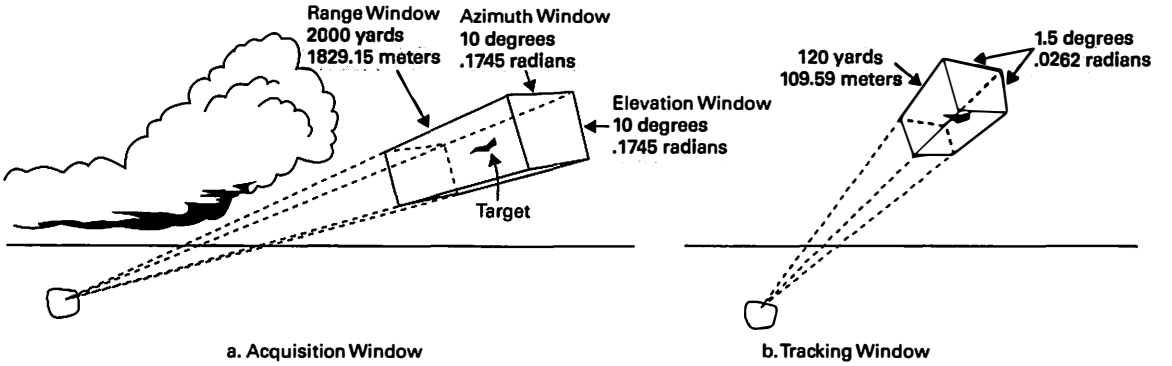


FIGURE 5-22 *Track-while-scan acquisition and tracking gates*

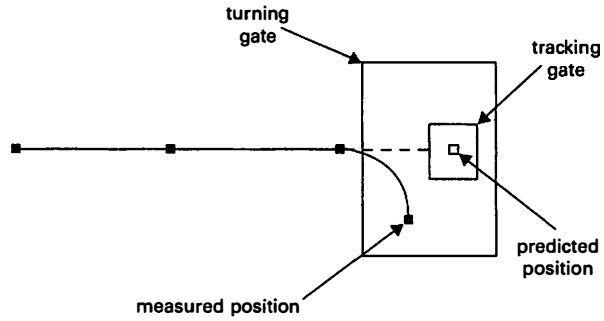


FIGURE 5-23 *Use of a turning gate to maintain a track on a maneuvering target*

generated until a smooth track is achieved. In the event that a valid return does not appear within the tracking gate on a subsequent scan, ADT will enter some type of turn-detection routine.

A common means of dealing with a turn or linear acceleration of the target is the turn-detection gate. The turn-detection gate is larger than the tracking gate and is co-located with it initially, employing separate logic and algorithms that are different from the tracking routine. The turn-detection gate size is determined by the maximum acceleration and turn characteristics of valid tracks; see figure 5-23 and figure 5-24.

Some systems will maintain one track with the original tracking gate and one with the turn-detection gate for one or more scans after initial turn detection, allowing for the possibility of inaccurate observation, clutter, or possible decoys prior to the determination of which is the valid track position. Only one of these tracks would be displayed to the operator, that one being determined by a programmed validity routine.

Track Correlation and Association

Target observation on each radar scan that survives hit-pattern recognition and clutter rejection functions is initially treated as new information prior to comparison with previously held data. Logic rules established in the digital computer program, which perform correlation and association of radar return with currently held tracks, are the key to preventing system saturation in high-density environments.

Generally speaking, target observations that fall within the boundary of the tracking gate are said to correlate with that track. Each observation is compared with all tracks in the computer's memory and may correlate with one track, several tracks, or no tracks. Conversely, a track gate may correlate with one observation, several observations, or no observations. Whenever there is other than a one-to-one correlation of observations with tracking gates, tracking ambiguity results; see figure 5-25.

Resolution of track ambiguity Track ambiguity arises when either multiple targets appear within a single track window or two or more gates overlap on a single target. This occurrence can cause the system to generate erroneous tracking data and ultimately lose the ability to maintain a meaningful track.

If the system is designed so that an operator initiates the track and monitors its progress, the solution is simply for the operator to cancel the erroneous track and initiate a new one. For automatic systems, software decision rules are established that will enable the program to maintain accurate track files, resolve ambiguity, and prevent saturation by false targets.

Depending upon the specific use of the system, many different sets of rules are employed to resolve this ambiguity, an example of which is outlined below:

1. If several video observations exist within one tracking gate, then the video closest to the center of the gate is accepted as valid.
2. If several tracking gates overlap one video observation, the video will be associated with the gate with the closet center.

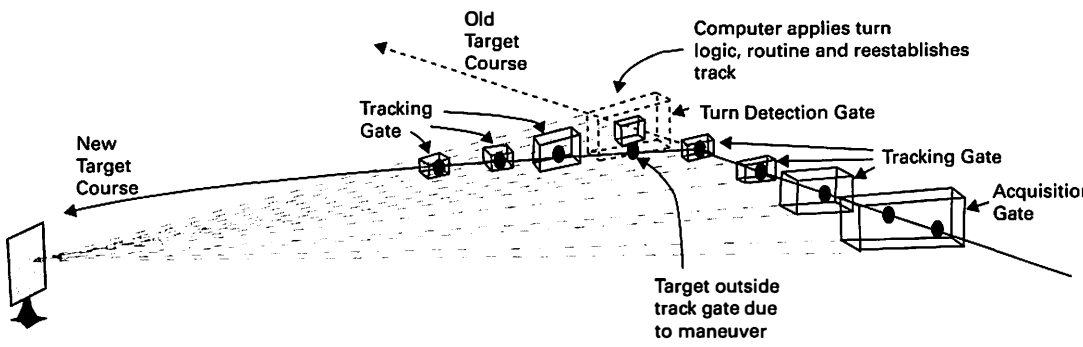


FIGURE 5-24 *TWS processing*

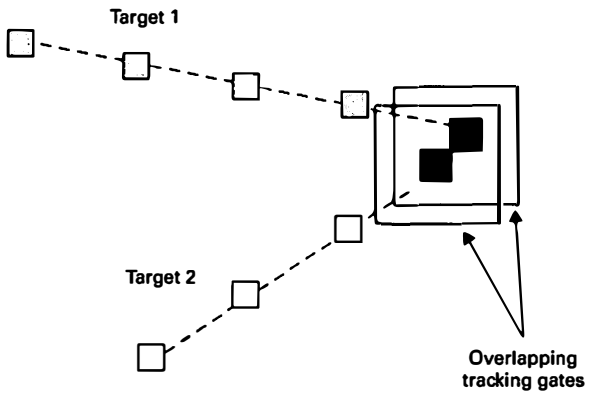


FIGURE 5-25 *Crossing tracks cause ambiguity that must be resolved.*

3. If two contacts cross paths, then the tracking gates will coast through one another, and recorrelation will occur when the gates no longer overlap.
4. If one video observation separates into two or more, rule 1 above is applied until there is clear separation in velocity and position and a new tracking gate is generated for the uncorrelated video. Rule 2 is then applied until the gates no longer overlap.

Track Initiation and Track File Generation

Concurrent with generation of the acquisition gate, a track file is generated in order to store the position and gate data for each track. In addition to the basic position and window data, calculated target velocities and accelerations are also stored in each track file. For ease of calculation and interchange of information with other systems, all data are converted from polar to rectangular coordinates by the computer. Track files are stored within the computer subsystems' memory in rectangular coordinates, and the data are used to perform the various calculations necessary to maintain the track.

Table 5-2 illustrates the data structure of a hypothetical target track file. Note that within the computer, position data have been converted from polar coordinates to rectangular coordinates. Each track file occupies a discrete position in the digital computer's memory. As data are needed for computation or new data are stored, the portion of memory that is allocated for the required data will be accessed by the system software.

In this manner, tactical data, in addition to the tracking data, may be stored in the "track" file—for example, electronic support data, engagement status, and IFF information. (*Note*: It is equally important to track friendly forces as well as hostile forces.) The generation of the track file begins with the initial storage of position data along with a code to indicate that an acquisition window has been established.

If target position data are obtained on subsequent scans of the radar, the file is updated with the coordinates, the velocities and accelerations are then computed and stored, and the acquisition gate code is canceled. The acquisition gate is then decreased in size relative to that of the tracking gate, and the track code is stored, which indicates an active track file.

As the radar continues to scan, each input of data is compared with the gate positions of active track files until the proper file is found and updated. The techniques of computer data file sorting and searching are beyond the scope of this text; however, it should be noted that the search for the proper track file is generally not a sequential one-to-one comparison. Such a method is much too slow to be used in a system where speed of operation is one of the primary goals.

TABLE 5-2 *Track file structure*

Track Number (Identifier)		
Position History		
X_n	Y_n	Z_n
X_{n-1}	Y_{n-1}	Z_{n-1}
X_{n-2}	Y_{n-2}	Z_{n-2}
Velocity History		
$\dot{X}_n$	$\dot{Y}_n$	$\dot{Z}_n$
$\dot{X}_{n-1}$	$\dot{Y}_{n-1}$	$\dot{Z}_{n-1}$
$\dot{X}_{n-2}$	$\dot{Y}_{n-2}$	$\dot{Z}_{n-2}$
Acceleration		
$\ddot{X}_n$	$\ddot{Y}_n$	$\ddot{Z}_n$
$\ddot{X}_{n-1}$	$\ddot{Y}_{n-1}$	$\ddot{Z}_{n-1}$
$\ddot{X}_{n-2}$	$\ddot{Y}_{n-2}$	$\ddot{Z}_{n-2}$
Window		
Acquisition code	Track code	
Position of gate center	Position of gate center	

Track Gate Prediction, Smoothing, and Positioning

As was discussed in the earlier sections on servo-controlled tracking systems, tracking errors were generated as a result of the target moving off the antenna axis. It was the task then of the error detectors and servo systems to reposition the antenna axis onto the target. During the

process of repositioning the antenna, the system response motion was smoothed by employing rate (velocity) and position feedback. Recall that this feedback was accomplished by electrical and mechanical means within the servo system. Recall also that in general these *systems lagged the target*—that is, the system responded to target movement.

In a track-while-scan system, tracking errors also exist due to target motion. The tracking gate now has replaced the "tracking antenna," and this gate must be positioned dynamically on the target in a manner similar to that of the "tracking antenna."

However, there is no "servo" system to reposition and smooth the tracking gate's motion. This repositioning and smoothing must be done mathematically within the TWS algorithm. To this end, smoothing and prediction equations are employed to calculate the changing position of the tracking window. Instead of the system "lagging" the target, *the tracking gate is made to "lead" the target,* and smoothing is accomplished by comparing predicted parameters with observed parameters and making adjustments based upon the errors derived from this comparison.

The classic method of smoothing track data is by the α-β tracker or α-β filter described below. This simple filter is ill-suited to extreme target maneuvers and in most current systems is increased in complexity to what is called the Kalman filter. Among other things, the Kalman filter is capable of dealing with higher-order derivatives of target motion (i.e., beyond acceleration).

In some systems, complete track file information is retained in the separate command and control system computer, with the TWS computer retaining only position data, velocity, and acceleration.

The Tracking Algorithm Systems Equations

Smooth position $$P_{sn} = P_{pn} + \alpha(P_n - P_{pn}) \quad (5\text{-}1)$$

Smooth velocity $$V_n = V_{n-1} + A_{n-1}T + \frac{\beta}{T}(P_n - P_{pn}) \quad (5\text{-}2)$$

Smooth acceleration $$A_n = A_{n-1} + \frac{\gamma}{T^2}(P_n - P_{pn}) \quad (5\text{-}3)$$

Predicted position $$P_{pn+1} = P_{sn} + V_nT + \frac{1}{2}A_nT^2 \quad (5\text{-}4)$$

(future position)

where:

P_n = target's position measured by the radar during scan n
P_{sn} = smoothed position after scan n
V_n = smoothed velocity after scan n
A_n = smoothed acceleration after scan n
P_{pn} = predicted target position for scan n
T = scan time (1 sec. in this case)
α, β, γ are system constants that determine the system response, damping, and gain, etc.

Note: P_n represents the three Cartesian coordinates.

Let us now examine equations 5-1 through 5-4 in greater detail, and in the process, correlate the track-while-scan functions to those of the servo tracking systems discussed earlier.

Equation 5-1

$$P_{sn} = P_{pn} + \alpha(P_n - P_{pn})$$

Equation 5-1 is analogous to the position feedback realized in the servo tracking systems by the radar error signal output. The subscript n refers to the current scan of the radar. The predicted position for the current scan P_{pn} is modified by the error between the observed position and the predicted position ($P_n - P_{pn}$); this yields a new, updated position, P_{sn}, for the scan. This updated position is then employed in equation 5-4 to calculate a future target position for the next scan ($n + 1$).

Equation 5-2

$$V_n = V_{n-1} + A_{n-1}T + \frac{\beta}{T}(P_n - P_{pn})$$

Equation 5-2 is analogous to the rate (velocity) feedback achieved by the tachometer loop in the servo tracking system. In equation 5-2 the target velocity is updated and modified by comparing the observed position with the predicted position and then dividing the position error by the time of the scan to obtain velocity $(P_n - P_{pn})/T$.

The velocity is further refined by modifying it by the acceleration $A_{n-1}T$.

Equation 5-3

$$A_n = A_{n-1} + \frac{\gamma}{T^2}(P_n - P_{pn})$$

Equation 5-3 develops the updated target acceleration to be employed in smoothing velocity on the next scan and in predicting the target position for the next scan. The old acceleration value is modified by the tracking error derived from the term $(P_n - P_{pn})/T^2$.

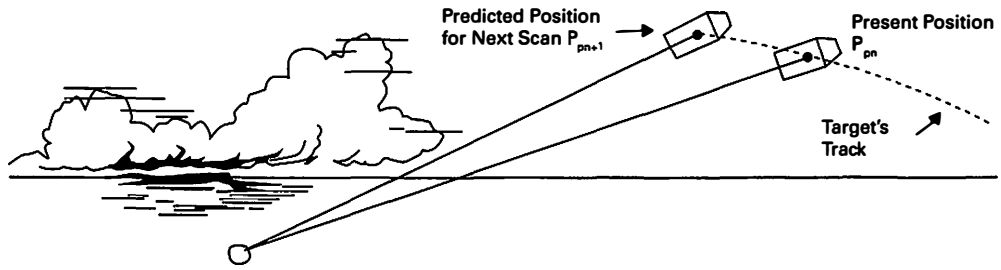


FIGURE 5-26 *Dynamic volumetric gate positioning*

Equation 5-4

$$P_{pn+1} = P_{sn} + V_n T + \frac{1}{2} A_n T^2$$

Equation 5-4 is simply the classic displacement equation derived in physics. It is used to predict future target position for the next scan; see figure 5-26. The predicted position for the next scan P_{pn+1} is obtained by modifying the smoothed present position P_{sn} by the velocity and acceleration obtained from equations 5-2 and 5-3.

Successive applications of this system of equations serve to minimize the tracking error in a manner similar to the repositioning of the "tracking antenna" in the servo tracking system. It should be noted at this point that although the techniques of tracking differ between the two (servo and TWS), the function and the concepts exhibit a direct correlation.

TWS System Operation

The annotated flow diagram of figure 5-27 takes the logical components of the TWS algorithm and combines them into a complete tracking algorithm. The reader should follow the flow through at least three radar scans in order to properly understand the system dynamics.

For the purpose of this illustration, assume that the system begins with a new detection and that tracking can be initiated on the second scan. In actual practice, three or more scans are usually necessary to establish a valid track. It should be observed that the entire algorithm is repeated for every target on every radar scan.

Active Tracking Radar

Conventional TWS procedures applicable to mechanically scanned radar systems do not fully exploit the potential of

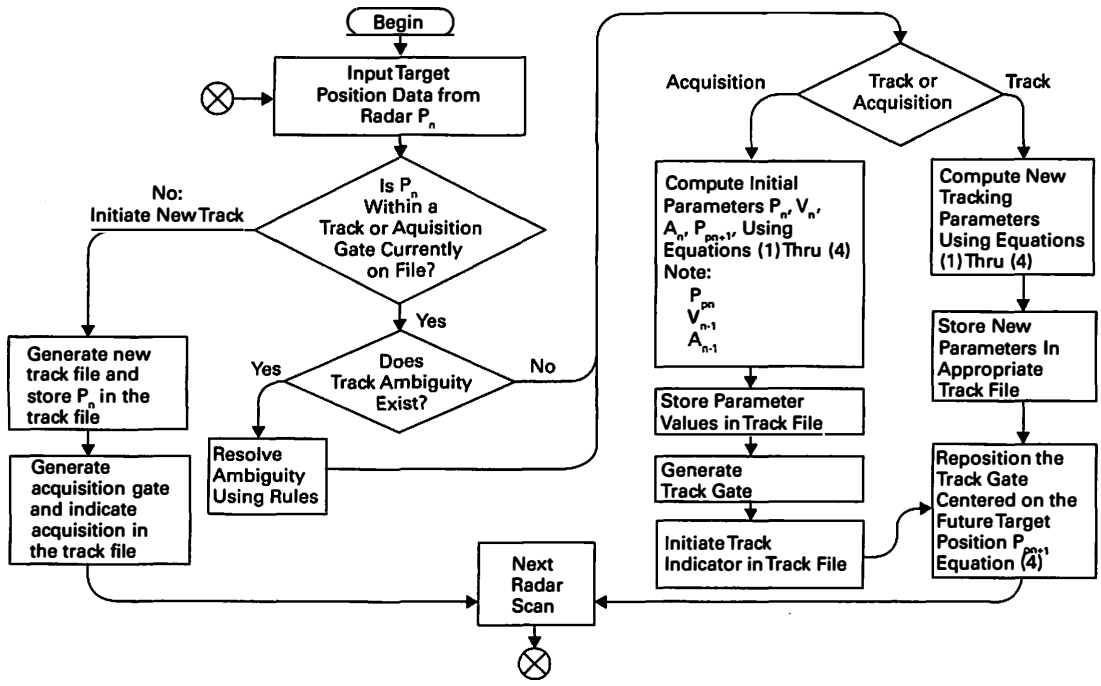


FIGURE 5-27 *Simplified TWS algorithm flow diagram*

radars employing electronically scanned arrays that steer beams in all axes. The ability of these systems to position beams virtually instantly and to conduct many types of scans simultaneously provides great flexibility in planning tracking strategies.

These radars conduct what amounts to a random search, employing different pulse widths and PRTs applicable to the situation. When a return that meets required parameters is observed, the radar immediately transmits several beams in the vicinity of the target, while continuing its search in the remainder of the surveillance area.

This results in the immediate establishment of a track without the relatively long (0.5 to 4 seconds) time period waiting for the antenna to return on the next scan, which is experienced in mechanically scanning radars. This technique, called active tracking, results in greatly decreased reaction time and more efficient use of the system.

Integrated Automatic Detection and Tracking (IADT)

By extending the concepts described in this chapter and adding additional computer capability, it is possible to combine the outputs of several radars that are co-located, as on ships or at shore-based command and control systems. Systems such as the AN/SYS-2 IADT develop a single track file based on the outputs of several radars. When multiple radars are employed with different scan rates, a separate external timing reference must be employed, which becomes the scan rate for the IADT and the synthetic video display.

Track updating and smoothing occur similarly to a TWS system. Observations are accepted from the radar having the oldest unprocessed position data sector by sector as the synthetic scan passes. This provides the first available position report, no matter which radar produced it. Thus, position reports are used as they occurred in real time, and no position report is accepted out of order.

The advantage of IADT is that it reduces loss of data due to individual radar propagation and beam characteristics while allowing quality weighting of data relative to its source. IADT systems can accept the output of TWS or non-TWS radars as well as that derived from IFF transponders.

Summary

This chapter has presented an analysis of automatic tracking systems. The tracking function was traced from the generation of position error magnitude and direction to the reposition of the antenna to reduce the error. Two methods of generating position error were investigated in detail, conical scan and monopulse.

Conical scan produces an error signal by causing the beam of sensor energy to be rotated rapidly about the sensor (antenna) axis.

Monopulse systems achieve the same result by electronically dividing the radiated energy into four separate quadrants, which form the beam. The relative amplitudes of the returning energy in the four quadrants are compared, and an error results when there is an imbalance in signal strength.

To improve the tracking of multiple targets and to lead vice lag the target, the track-while-scan concept was developed. The central idea underlying any TWS system is that the sensor itself continues to perform its primary function of search (scanning) and data input while the remainder of the system performs the target tracking function. The sensor simply provides target position data to the computer subsystem where target velocities and position prediction are calculated.

In a military application, the major advantage of a TWS system is the elimination of the process of target designation from a search radar to a fire control radar. Tracking information developed in the TWS system is used as a direct data input to the computation of a fire control solution. Therefore, as soon as a target is detected, a fire control solution is available without the inherent delay caused by the designation process. Time required from first detection to fire control solution is on the order of seconds for a TWS system, as opposed to tens of seconds or even minutes for a manually designated system employing separate search and fire control sensors.

---

[i] Hall, 125

[ii] Hall, 126

[iii] Hall, 126

[iv] Hall, 127

[v] Hall, 127

[vi] Hall, 130

[vii] Hall, 131

[viii] Hall, 131

6 Electronic Combat

Introduction

Electronic combat (EC) started as early as World War I when Germany produced a device that detected the electromagnetic field variations of telegraph wires. This gave the Germans the ability to eavesdrop on the allies' communications and to exploit the use of the electromagnetic spectrum.

Technological advances over the last century have rapidly accelerated the development of this facet of modern warfare to encompass the entire magnetic spectrum. Electronic combat continues to be vital, influencing both tactical and strategic decisions, and includes all those means of exploiting an adversary's own activity as a means of determining its intentions or reducing its effectiveness. This involves using energy radiated as a result of an adversary's presence or activity as a means of detection and identification of potential weapon systems and platforms. In addition, weakness in sensors, fuzes, computer processing, and personnel training are exploited to deceive an adversary as far as the actual tactical situation or to cause malfunction of an adversary's equipment.

Electronic combat applies to the entire spectra or waves—from sound waves to electromagnetic waves of any frequency. Since the most prevalent use of the electromagnetic spectrum is for weapons, sensors, and communications in the radio frequency—that is, several KHz to tens of GHz—most of electronic combat is focused in this region. Infrared (IR) and acoustic portions of EC are also of interest, and these countermeasures will be addressed specifically in chapters 8 and 11.

Avoiding detection and identification by an enemy sensor or weapon is also part of electronic combat. This includes the concept of passive defense measures to reduce the enemy's effectiveness and to hamper destructive systems.[i]

A major difficulty in studying countermeasures is how much the subject involves classified information. Detection depends on sensors, which ultimately depend on a receiver to determine the presence or absence of a contact. Possessing the technical means to disrupt or deceive that receiver is an advantage to be guarded very closely. Hence, this discussion of countermeasures will be generic, emphasizing the principal areas in an expanding field.[ii]

Electronic Combat

Electronic combat is considered any military action involving the use of electromagnetic energy to determine, exploit, reduce, or prevent an enemy's use of the electromagnetic (EM) spectrum, and the actions that retain friendly use of the EM spectrum. EC has three branches:

1. *Electronic support* (ES)—Passive surveillance of the EM spectrum to detect the enemy's position, strength, and intention, and warnings of targeting or homing.
2. *Electronic attack* (EA)—Preventing or reducing the enemy's use of the EM spectrum and promoting uncertainty.
3. *Electronic protection* (EP)—Protecting friendly combat capabilities against the undesirable effects of enemy electronic attack. This is accomplished by considering radar operating parameters, signal-processing techniques, and design philosophy.

Basic Electronic Combat Receivers

The challenge in electronic combat is to obtain the raw data, which in turn is analyzed to determine the enemy threat. To obtain this data, the collection station will

require dedicated receiving equipment. The electronic support (ES) receiver differs from an ordinary radio wave receiver both in its essential design and in the auxiliary equipment associated with it. The necessary requirements for an ES receiver are:

Wide spectrum surveillance (wide bandwidth capability) The frequency of an enemy radar is not known beforehand. This means that the frequency spectrum must be searched from roughly 30 KHz to 50 GHz. This range is too large for one receiver, so the solution has been to connect many different frequency-sensitive circuits called tuners to a common chain of main amplifiers.

Wide dynamic range Since the operating distance from a radar is often unknown, the receiver must be able to receive and process both very weak and very strong signals without changing the signals characteristics for analysis.

Unwanted signal rejection (narrow bandpass) Many other signals will exist, with frequencies close to the signal of interest. The receiver should discriminate well between the frequency it is tuned to and signals at other frequencies.

Angle-of-arrival measurement capability Precise bearing measurements allow for locating the transmitter by taking bearings at different locations and then plotting these bearings on a chart to locate the transmitter by triangulation. An airborne or ground-based digital computer can also be programmed to perform the same by employing special target motion analysis (TMA) techniques.

Signal analysis capability This provides a means of determining the signal modulation, sidebands, pulse width, and PRF. From this information, the signal can be identified and associated with a specific threat or platform. This is most efficiently performed by a digital computer, but can be done by manual analysis and consulting publications.

Display The type of display is dictated by the way in which the receiver is used. This varies between the "Fuzzbuster"-type warning light and audio tone devices of Vietnam-era fighter aircraft detection systems to very complex signal-analysis scopes and computer-controlled alphanumeric displays.

Recording system There is great intelligence value in electronic emissions of all types, including commercial television, radio, and computer data. For this reason ships, aircraft, submarines, and marine units are equipped with magnetic tape and other devices to record emissions for further analysis at major intelligence centers.

Before discussing specific types of receivers, it is useful to develop a "common language." The following terms will be used concerning ES receivers.

- *Bandpass*—Range of frequencies capable of being detected by a receiver
- *Receiver Sensitivity*—A measure of minimum power in a signal, present at the receiver, that can be detected by that receiver
- *Bandpass Filter*—An electronic device designed to pass only a certain range of frequencies through its circuitry
- *Probability of Intercept (POI)*—Probability that a signal will be detected: usually a function of the sensitivity and bandpass of a receiver
- *Signal Analysis*—Process of categorizing a received signal by examining it through manual or electronic means
- *False-Alarm Rate*—Occurrence when a receiver misidentifies a signal
- *Parametric Data*—Characteristics of a received signal, generally RF, PW, PRF, PRT, scan rate, and scan type

Direct Detection Receiver

A direct detection receiver is the simplest of ES receivers. It is designed to indicate the presence of a signal within its sensitivity and bandpass. This receiver utilizes either single or multiple bandpass filters in order to determine the frequency of a detected signal. Direct detection receivers are very reliable, small in size, lightweight, and relatively cheap; see figure 6-1. It is tactically significant that these

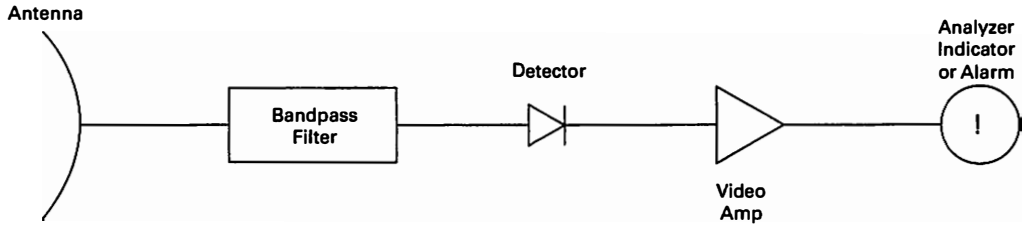


FIGURE 6-1 *Direct detection receiver*

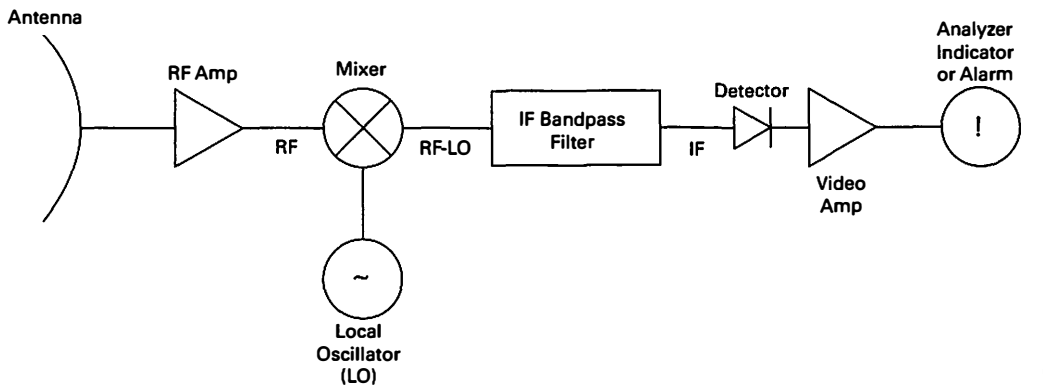


FIGURE 6-2
Superheterodyne receiver

receivers be designed to indicate the presence of a signal within a fairly narrow bandpass, for instance, the range of frequencies where a weapon control radar would be expected to be operating.

In its simplest form, a direct detection receiver would "alarm" when it detected a signal whose frequency lies within its bandpass. In practice, these receivers have fundamental signal-analysis circuitry that allows them to analyze other parameters of an intercepted signal in addition to frequency. This more detailed analysis reduces the false-alarm rate of the receiver. As such, direct detection receivers have a very high probability of intercept (POI) within their bandpass.

Radar warning receivers (RWR) used in tactical aircraft are excellent examples of direct detection receivers. RWRs display no parametric data of the intercepted signal. They do, however, indicate the presence of, and relative bearing to, an emission that has been predetermined hostile based on parametric data.

The principal problem with direct detection receivers is poor performance in dense electromagnetic environments. The same wide bandpass that allows detection of signals within a broad range of frequencies prevents the discrimination between signals within the bandpass and allows an excessive amount of noise to enter the receiver.

One way to improve frequency discrimination and reduce receiver noise is to narrow the bandpass. Unfortunately, narrowing the bandpass also reduces the range of frequencies that a receiver can detect. Certain types of direct detection receivers employ a bank of narrowband bandpass filters in order to improve their performance. An instantaneous frequency measurement (IFM) receiver is a type of electronic intelligence receiver that uses this technology.

Superheterodyne Receivers

The most common type of receiver used for signal intelligence gathering is the superheterodyne receiver; see figure 6-2. In a simple superheterodyne receiver, a signal entering the device is mixed with the output of a tunable local oscillator (LO). The mixing process is called heterodyning and produces a sum frequency and a difference frequency. Processing the difference frequency is called superheterodyning.

A narrowband bandpass filter tuned to a specific intermediate frequency (IF) follows the mixer in the receiver circuitry. When the local oscillator is tuned through a frequency such that the difference between the signal RF and the LO frequency equals the IF frequency, signal will pass through the bandpass filter and be amplified before reaching a detector. The signal can then be displayed for analysis, its frequency being determined by the position of the LO at the time IF is achieved.

In actual use, superheterodyne receivers are somewhat more complicated. The receivers can be separated into channels by using multiple LOs and IF sections, sometimes in parallel and at other times in series. This technique allows for faster, more accurate scanning of the receiver's bandpass.

The fact that a superheterodyne receiver passes a smaller range of frequencies for a fixed bandwidth percentage is called arithmetic selectivity and is one of its major advantages over conventional receivers.[iii] This means a superheterodyne receiver is more selective and has a smaller band of frequencies that it will accept, process, and amplify for the same bandwidth percentage compared to a conventional receiver.

The range of frequencies over which the local oscillator is tuned determines the receiver's bandpass, and width of the IF bandpass filter determines its frequency "accuracy."

A larger band of the microwave spectrum can be scanned, and determination of the signal frequency can be made with greater accuracy. In addition, the narrow IF filter and amplifier allow amplification and detection of more signals with less noise, thereby improving receiver sensitivity and performance.

Unfortunately, superheterodyne receivers are not without disadvantages. They tend to have higher false-alarm rates, particularly where strong signals are concerned because of signal harmonics. They also have a lower probability of intercept than direct detection receivers, because a superheterodyne receiver "scans" through its entire bandpass, looking at an "IF bandpass size" of the spectrum at a time rather than staring at the entire bandpass. Despite these limitations, most operational and strategic electronic intelligence systems utilize some type of superheterodyne receiver.

Electronic Attack

The second major division of electronic combat is electronic attack (EA)—the prevention or reduction of the enemy's use of the EM spectrum. Of the three divisions of electronic combat, EA is probably the best known, because EA tends to be visualized as "black boxes" that jam or deceive the enemy.

Often it appears that if one understands the black boxes, then there is an understanding of EA; but such an attitude is very narrow because it ignores all other types of EA except jamming and deception. Thus, the approach in this section is more general; an attempt is made to lay down the framework within which the black boxes function and show other types of EA.

Electronic attack may be employed against enemy sensors and communication systems. The emphasis of EA employment in this section is against enemy sensor systems because they have an immediate threat to friendly forces. However, some mention of the theory and practice of employing EA against communications systems is considered appropriate, particularly in the contemporary Navy, which is heavily dependent upon communications, including the various computer data links that provide the backbone to Fleetwide and Joint command and control efforts.

From a strategic point of view, using EA against enemy communications must be balanced with losing the opportunity to gain valuable information by eavesdropping on the same communications. Tactically, however, it may be very advantageous to jam the enemy communications system in order to cause a breakdown in the battle plan.

This was vividly illustrated during the 1973 Middle East War when the Egyptians successfully jammed the Israeli UHF/VHF radio frequencies, resulting in a complete disruption of the Israelis' air-to-ground communications and consequently a significant reduction in the effectiveness of their close air support.

Typically, EA is used against a wider array of emitters to include: long-range passive detectors; radar-warning picket ships; airborne radar patrols (AWACS); long-range early warning radar sets; ground-controlled intercept radar sets; fighter intercept radar; missiles guided by radar; radio and radar navigation equipment; electronic bombing equipment; electronic identification equipment (IFF); terrain-following radar; antiaircraft artillery (AAA); fire control radar; and surface-to-air (SAM) control radar, and so on. The particular method used will depend upon the tactical situation.

Basic Principles of EA Effectiveness

The basic purpose of EA is to interfere with the operation of air and surface defense system sensors in order to reduce the information content of the signals and to induce errors. When conducting EA, it is often difficult for an EA operator to determine if the jamming techniques being employed are effective. To understand this difficulty, it is necessary to look at two different types of radar systems, monotrack and track-while-scan (TWS).

A monotrack system follows only one target at a time and may use a dedicated tracking radar, such as a conical scan or monopulse radar, to align the axis of the antenna with the target. The system then measures the azimuth and elevation angles and combines that information with the radar range to determine target position. Over a period of time, the system can then measure both velocity and acceleration.

As described, monotrack systems can follow only one target at a time, and the antenna must be pointed at the target for the system to track. EA can deny range to the radar, but the antenna can still seek out the direction of the received EA and thus determine the direction of the target. In addition, certain inverse gain techniques cause the antenna not to point at the target, providing the jammed system false information. This also gives the EA operator a direct measure of effectiveness because the antenna is no longer pointing at the target.

Now consider the TWS system, which tracks more than one target. It does so by looking at targets in sequence by covering the total area of interest and noting where targets are detected. For each detection it stores position information while continuing its normal scan and compares

that with the position information on the next scan. After two or three scans, TWS systems are able to tell the position of each target and generate accurate tracks. Consequently for the EA operator, the TWS system observes only the EA platform periodically, providing no direct measure whether it is tracking the EA platform, the screen platform, or both.

It is important to stress that EA does not have to completely prevent tracking to be effective. Delaying the establishment of a solid track on a target, causing a moment of confusion, or forcing the decision maker to wait just those few more seconds to be sure of the proper response can enable weapons to penetrate an adversary's defenses.

The Four Classes of EA

In general there are four fundamental classes of electronic attack. The first method uses active radiators and is commonly referred to as jamming. Any radar, data link, or communication system that emits electromagnetic radiation in the microwave portion of the spectrum can be jammed. There are many different techniques that can be used to jam enemy electromagnetic emissions. This text will separate these techniques into two broad categories: denial jamming and deception jamming.

The second method uses medium modifiers that change the electrical properties of the medium between the targeted platform and the radar, including techniques that reflect the radar energy.

A third method includes applying radar-absorbing materials to the screened platform and both electronic and mechanical echo (BLIP) enhancers for decoys.

The fourth method is physical destruction and uses blast or fragmentation warheads to physically damage equipment.

Transmitters

Before discussing jamming, the specifics of EA transmitters must be addressed. Effectiveness of an EA transmitter (jammer) depends upon power output of the transmitter, losses in transmission, antenna gain in the direction of the victim receiver, and transmitter bandwidth. EA success is also based on the victim's radar receiver and is a function of the victim's bandwidth, antenna gain, and radar cross-sectional area of the target being masked. In order to be effective, the EA transmitter must be capable of emitting enough power in the bandwidth of the victim receiver to mask (jam) its intended signal or to simulate a deceptive signal realistically.

In order to meet these requirements, most EA transmitters are designed to be versatile in operation. When the EA transmitter is used against only one missile, one radar, or one communications device, or against a few such devices grouped closely in frequency, the transmitter can concentrate its available output power into a narrow spectrum.

On the other hand, if the EA transmitter must operate against several devices separated in frequency, it must spread its available power output over a correspondingly increased spectrum. For example, a 1,000-watt transmitter that emits its energy in a 10-MHz spectrum is developing a power density of 100 watts per MHz. If the same transmitter must proportion its energy to cover a spread of 100 MHz, its power density is therefore 10 watts per MHz.

As indicated previously, the effectiveness of the EA transmitter depends upon its power density within the bandwidth of the victim receiver, after considering antenna factors and propagation path losses.

Denial Jamming

Denial jammers are intended to overload the opposition's receiver so that its use is denied to the enemy. The jammer creates a noise signal powerful enough to mask the signal the receiver is trying to detect. In terms of signal-to-noise ratio, the jammer raises the noise in the front end of the receiver to the point where the signal power is less than the minimum signal power required for detection. As such, these jammers rely upon pure power and brute force to degrade enemy signal reception; see figure 6-3.

Denial jammers are essentially noise jammers. Noise jamming is simple in concept. All radars are susceptible

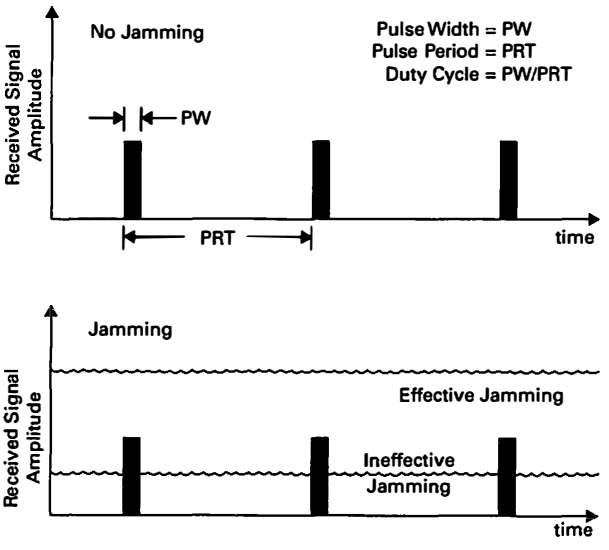


FIGURE 6-3 *Effective and ineffective jamming showing burnthrough*

to the effect of "natural" sources of white Gaussian (random) noise in the atmosphere. A noise jammer utilizes a noise diode or other random noise source to modulate a high-powered carrier. The noise-modulated carrier is then amplified to a high-power level and pushed out of a high-gain antenna.

When the target radar has a transmitter and receiver, jammers, particularly denial jammers, have an inherent power advantage over victim receivers. This is because the jamming signal has to propagate in only one direction. As such, the noise power generated by a jammer has only half the atmospheric attenuation and spreading losses that the target radar signal experiences.

Although the primary victims of denial jammers are acquisition radars, they can be used against any number of targets such as communications systems, navigation systems, and fuzing systems.

Similarly to other weapon systems studied, denial jammers have advantages and disadvantages relative to deception techniques. An advantage to this type of jamming is that it is relatively simple to implement. However, disadvantages include: its effectiveness is limited by maximum radiated power; countermeasures are relatively simple and widely employed; and finally these jammers are vulnerable to home-on jam-type missile systems.

Burnthrough is the single most important concept to grasp for tactical employment of a denial jammer. Burnthrough is usually defined as the range at which the strength of the radar echo becomes equal to or greater than the EA signal. To state it another way, burnthrough is the range that the jammer is no longer putting out enough power (noise) to overcome the signal power that is trying to be masked; see figure 6-4.

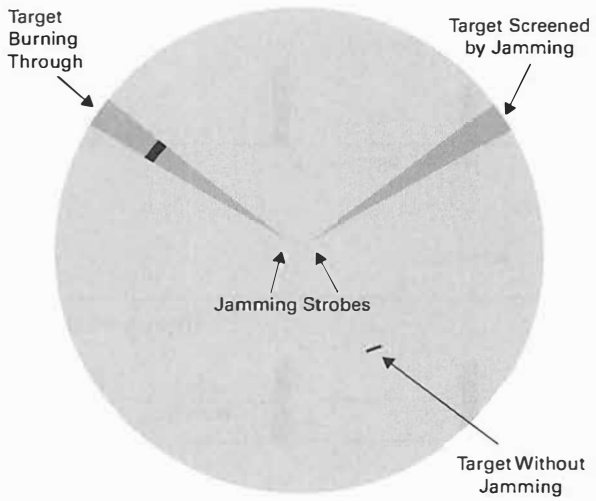


FIGURE 6-4 *Radar PPI scope with jamming*

To illustrate the point, assume that the gain of the targeted receiver and power output of the jammer are fixed. By making this assumption, the range at which an enemy receiver may "burnthrough" denial jamming can be calculated for two different scenarios.

Self-screening or escort jamming In this scenario the jammer is either masking its own platform or masking other platforms in the near vicinity, defined as within the jammer's beamwidth. This occurs normally when a unit carries jamming equipment for its own protection.

A trade-off results in this configuration between weight and space reserved for EA equipment and that reserved for sensors, weapons, and fuel. The trade-off is most critical in aircraft and least critical in ships. The self-screening method requires maintaining efficient jamming geometry between the victim radar and the jammer because the jammer and victim radar are always along the same line.

Burnthrough for this type of jamming can be calculated using the following equation:

$$R_{\text{burnthru}} = \sqrt{\frac{P_{\text{peak}} G_r C \sigma}{4\pi P_{\text{BW}} G_j B}} \qquad (6\text{-}1)$$

where:

P_{peak} = radar power output (w)
G_r = radar directive gain (dimensionless)
σ = target radar cross section (m^2)
R = radar to target range (m)
P_{BW} = jammer power per noise bandwidth (w/MHz)
G_j = jammer antenna gain (dimensionless)
B = radar receiver noise bandwidth (MHz)
C = minimum J/S ratio for target obscuration (dimensionless)

Stand-off jamming In this scenario the jammer positions itself along the axis of attack, but does not escort other units in the strike package. A unit performing stand-off jamming remains just outside the range of enemy weapons, providing screening for attacking units.

The advantage of standing off is that the jammer is safe from enemy home-on jam (HOJ) weapons, a submode of nearly all radio frequency homing weapons. The disadvantage of this geometry is that the attacking units become more vulnerable because the jammer must remain at longer range while the attacking units close the enemy.

Another less common type of off platform jamming is called stand-forward jamming. Stand-forward jamming occurs when the jamming unit is placed between the enemy sensors and the attack units. Although maintaining proper geometry between victim sensors, attacking units,

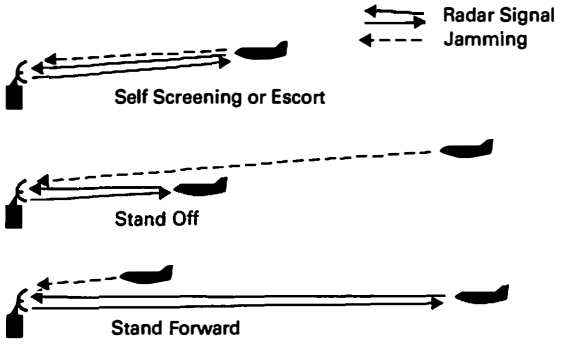


FIGURE 6-5 *Jamming tactics*

and the jammer is difficult, this method allows the most efficient use of jammer power by reducing spreading and attenuation losses. This situation is the most dangerous for the jamming unit because it is a prime target for all weapon systems and well within the capabilities of HOJ and antiradiation (ARM) weapons.

For stand-off jamming, the burnthrough range for the targeted platform is calculated using the following equation:

$$R_{\text{burnthru}} = \sqrt[4]{\frac{P_{\text{peak}} G_r C \sigma R_j^2}{4\pi P_{\text{BW}} B G_j}} \qquad (6\text{-}2)$$

where:

R_j = range from radar to jammer

The problem with denial jamming is that it's been around for a very long time, certainly long enough for many techniques to have been developed to counter its effect. Also, because denial jamming radiates 100% of the time, it is exceptionally hazardous to employ denial jamming against missile systems that employ home-on jam or track-on jam techniques, because the jamming signal acts as a beacon for these types of systems.

Within the general class of jamming, there are three different techniques for generating the noiselike signal to be used. The first is spot jamming where all the power output of the jammer is concentrated in a very narrow bandwidth, ideally identical to that of the radar. The other two are barrage and sweep jamming, which spread their energy over a bandwidth much wider than that of the radar signal.

Thus, spot jamming is usually directed against a specific radar and requires a parametric receiver to match the jamming signal to the radar signal. The other two techniques, however, can be used against any number of radars and require only a receiver to tell them that there is a radar present.

The difference between barrage and sweep jamming lies in the modulation techniques and size of the frequency band covered. Barrage jamming often uses an amplitude-modulated signal covering a 10% frequency band (bandwidth equal to 10% of the center frequency). Sweep jamming often uses a frequency-modulated signal, and the frequency is swept back and forth over a very wide bandwidth. Figure 6-6 illustrates these three types of jamming.

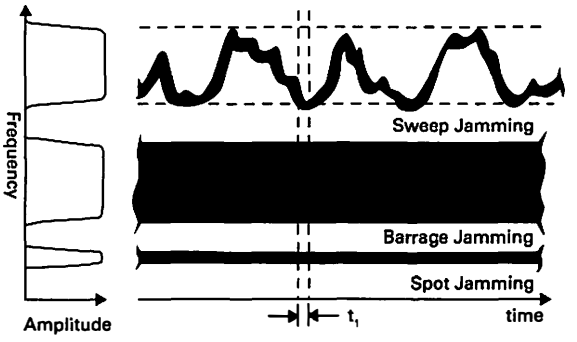


FIGURE 6-6 *Spot, barrage, and sweep denial jamming*

It is nearly impossible to match a jammer frequency exactly to that of a radiating radar; therefore, it is usually necessary to broaden the bandwidth of the noise so that it is greater than the radar bandwidth. A barrage jammer has a very wide bandwidth to cover all radars with frequencies in that band, whereas the spot jammer attempts to match as closely as possible a particular radar frequency.

This broadening of the jammer bandwidth causes the jammer to require more power than when it is exactly matched, because the power that matters for any radar is the power that is accepted by the receiver. This fact is usually accounted for by specifying the spectral power density that a jammer must have to jam a radar. Figure 6-7 illustrates the idea of power density by showing that a jammer

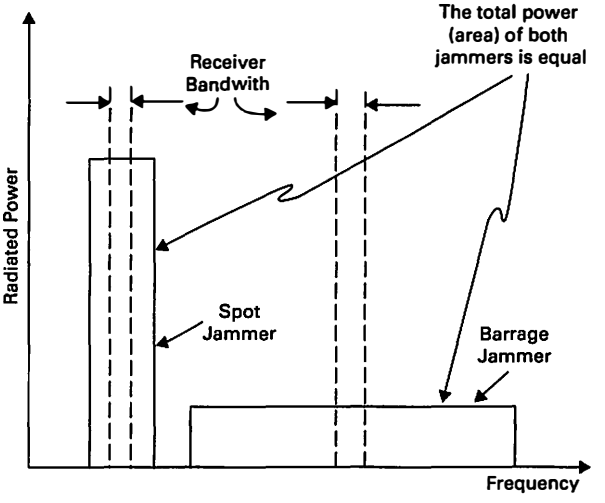


FIGURE 6-7 *The effect of bandwidth on jammer spectral power density*

of a given total power is more effective if its bandwidth is decreased. The usual means of specifying jammer power density is in watts per megahertz, w/MHz.

Since aircraft are limited in the total amount of jammer power they can carry, it is advantageous for the air defense network to use as many widely different frequencies for its radars as possible. This concept is usually called frequency diversity, and it forces the jamming platform to either carry a large number of spot jammers or spread their barrage and sweep jammer power in order to cover all the radars. Frequency diversity also eliminates the mutual interference of integrated forces during large operations. The ability of a single radar to change frequency to counter a spot jammer is called frequency agility.

Deception Jamming

To obtain a fire control solution there is a need to be able to track a target in azimuth, elevation, and range. Tracking the position of a target over time by observing these three parameters provides estimates of target location at some time in the future, something very useful in directing a weapon toward a collision.

Denial jamming "denies" the victim radar system this range information as previously mentioned. Unfortunately for the jammer, modern electronic counter-countermeasures such as automatic gain control (AGC) can reduce noise jamming effectiveness to the point where an accurate estimate of bearing is possible. In spite of the fact that range information is unavailable, if the missile fired has sufficient range and velocity (relative to the target), it may very well still hit its target, albeit at a significantly less-than-optimal intercept profile.

This is where deception jamming is used. Whereas denial jamming masks the location of the target with noise, deception jamming misleads or "spoofs" the radar by providing it with a false target location. Consequently, deception jamming is generally accomplished by devices called repeaters and transponders, and is sometimes also referred to as repeater jamming.

The theory of repeater operation is straightforward. However, actual implementation requires sophisticated circuitry. Basically, the radar signal is received, delayed, amplified, modulated, and retransmitted back to the radar.

A transponder differs slightly in that it plays back a stored replica of the radar signal after it is triggered by the radar. The transmitted signal is made to resemble the radar signal as closely as possible. Delay may be employed, but amplification is usually not used. An important point is that power requirements for a deception-type repeater are much lower than those for a noise jammer, since the repeater emits energy in pulses similar to the radar pulses. Its duty cycle is also similar to that of the radar.

By learning enough of the operating parameters of the victim radar through ES, weaknesses of the system can be exploited. Most deception jamming techniques take advantage of one or two parameters. Against acquisition-type systems, deception jamming attacks the PRF of the victim radar and against tracking-type systems, deception jamming attacks the scan rate.

As discussed earlier in chapter 3, a radar operating at a frequency of 9 GHz does not simply transmit at 9 GHz. The transmitted pulse has some finite bandwidth, which can be estimated using the following equation:

$$BW_{-3dB} = \frac{1}{PW} \tag{6-3}$$

where:

BW_{-3dB} = bandwidth of the receiver at the half-power point

PW = width of the transmitted pulse

The operators of denial jammers are interested in this parameter instead. If the victim radar has a bandwidth of 5 MHz, the jamming spot must be at least 5 MHz wide in order to be effective.

Operators of deception jammers are also interested in this parameter because it defines the spectrum of the transmitted jamming pulse. From the study of advanced radars, it is the spectrum of the reflected pulse that is analyzed for range information on the target.

False targets can therefore be generated by careful selection of the spectrum of the transmitted jamming pulse followed by transmitting this pulse at either the PRF or a multiple of the PRF of the victim radar. This allows the victim radar to detect returns at frequencies within its bandpass, which to the radar will appear to be targets.

Transmitting pulses at the PRF rate of the victim radar results in false targets at approximately the range of the real target. By varying the rate of these transmitted pulses by multiples of the actual victim radar's PRF, false targets are generated that surround the real target at both long and short ranges.

Tracking radars use sophisticated scanning techniques to keep a target within its boresight. In order to keep a target boresighted, tracking radars must scan the environment and continuously compare the position and strength of a return to some reference value.

The reference value is generally the scan frequency of the radar. The tracking system will compare this value to the frequency at which the target appears within its field of view. The tracking logic within the system uses

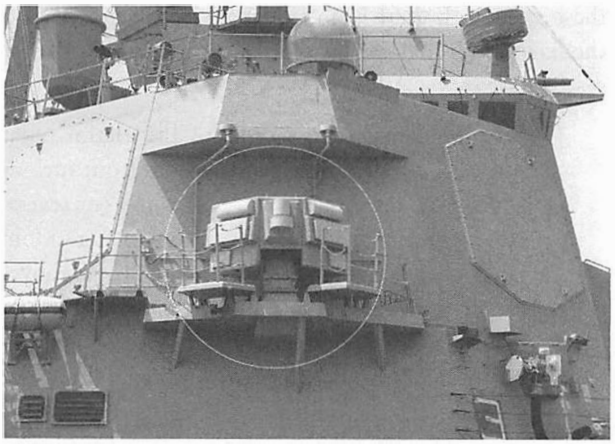


FIGURE 6-8 *Shipboard ES and EA antennas*

algorithms to generate aiming commands, which then drive elevation and bearing servo motors as described in chapter 5. The logic in the algorithms generally seeks to zero the difference between the reference and return frequencies, which will occur only when the target is boresighted by the antenna. More complex systems also take signal strength and phase into account as well.

Range deception If a repeater were to simply retransmit the received pulse as soon as it was received, it would reinforce the return echo and help rather than frustrate the victim radar. But if the received pulse (as opposed to the echo that returns to the radar) could be briefly stored and then transmitted a short time interval later, the radar would first receive the weak natural echo return followed by an identical but stronger pulse sent from the repeater. If a repeater transmitted a series of time-displaced pulses identical to the radar pulse, it could then produce a series of spurious targets, each at different ranges.

In automatic-tracking radars, the first step in the process of locking onto a target is for the operator to designate the specific target of interest by positioning the range-tracking gate over the target return video. Once this is done, the radar's receiver is, in effect, turned off until such time as an echo return is expected at the approximate range of the designated target, thereby making allowance for the velocity of the target.

This allows the deception repeater to operate in the "range-gate" or track-breaking mode. Initially, the repeater simply repeats back the received radar pulse without any delay, to allow for the radar's automatic gain control to adjust to the stronger signal, which it assumes to be the designated target.

Then the deception repeater begins to introduce increasing amounts of time delay—see figure 6-9—before retransmitting back the received radar pulse. Thus, the range-gate circuitry in the radar tracks the stronger pulse and gradually "walks off" from the true target range and makes the target appear to be at a greater range than it really is.

Similarly, the target can be made to appear at a closer range by delaying the received radar pulse long enough that it can be retransmitted back prior to receiving the next radar pulse. Then the deception pulse will arrive at the radar before the real echo pulse, producing a false target at a closer range.

This false target range information can cause significant aiming and guidance errors for antiaircraft guns and missiles that require command guidance from ground-based radars.

The simplest remedy against this type of jamming is to have tracking radar operators switch to a manual mode of operation. This remedy is effective because an operator watching a radar display can see the cover pulse move away from the aircraft return, and can therefore track the aircraft.

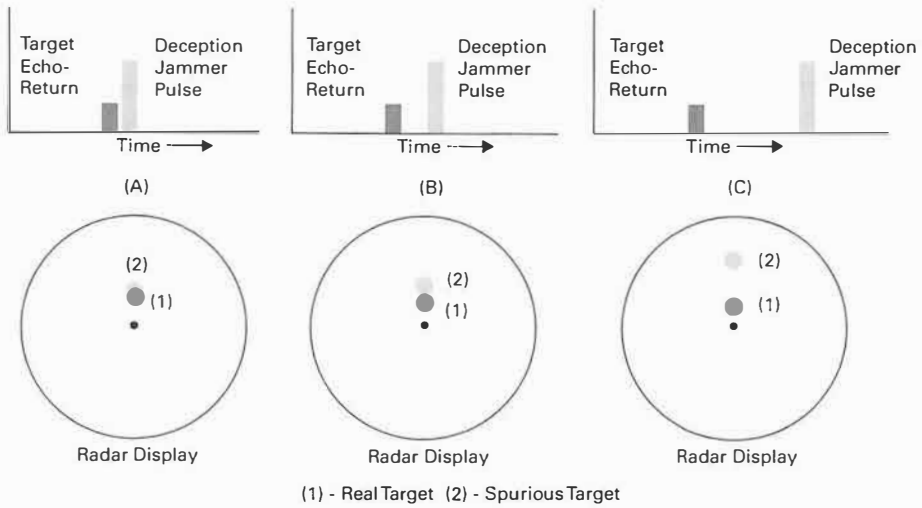


FIGURE 6-9 *Range deception as it would appear on an air-surveillance PPI*

Even though manual tracking will largely counter a repeater, manual tracking is never as smooth as automatic tracking. Thus, the weapon miss distance will increase, thereby increasing the probability of aircraft survival.

Angle deception Radar and command control systems can be confused by causing the radar to generate incorrect target bearing and elevation information. For this to be successful, the deception device must cause the radar to indicate the presence of a target at some time other than when the radar is at the target's bearing and elevation. There are two primary methods of achieving this effect.

The first is sidelobe angle deception. In this method, the sidelobes in the antenna radiation pattern must be evident to the EA unit. A false target pulse is then transmitted while the EA unit is at the bearing of a sidelobe of the victim radar. The radar circuitry is designed to register target angular position in the main lobe only, and therefore displays target video with an angular error equal to the angular displacement between the main lobe and the sidelobe involved.

This technique can be applied to any radar with ineffective sidelobe suppression or cancellation. By combining this method with range deception, many false targets at different ranges and bearings can be generated, causing confusion over the entire search volume of the victim radar, with much less required average power than equivalent noise jamming.

The second method is called angle tracking circuit deception. Monotrack radar such as that employing monopulse, conical scan, COSRO, LORO, and track-while-scan or active tracking radars can be deceived by causing their angle-sensitive circuitry to drive in some direction other than that which would correct the existing angular error at any given time. In doing so, the EA unit will cause errors in solution of the fire control problem and thus induce considerable error and/or delay in weapon employment.

Deception of angle tracking circuits sometimes involves specific equipment tailored to each angle tracking technique. For instance, the inverse conical scan (inverse gain) method is effective only against the conical scan tracker and does not work with the others mentioned above.

For all angle tracking methods, the EA unit must have knowledge of the scanning and tracking techniques employed by the victim radar. For the conical scan and lobe-switching radar, this may be obtained by monitoring the radar with an ES receiver. Monopulse, COSRO, LORO, TWS, and active tracking radars reveal nothing to an ES operator concerning their tracking methodology.

Therefore, good intelligence data are required to build deception devices and properly employ them against these systems. If there is not adequate intelligence, two of the techniques that can be employed against all of these systems are:

1. *Blinking.* Noise or a sample of the victim radar pulse is amplified and retransmitted from various widely separated points on the EA unit (or several closely spaced cooperating units) in a random fashion, causing enhancement of the usual movement (wander) of the point on the EA unit that the radar tracks. For smooth tracking and accurate solution of the fire control problem, the radar should track the centroid of the target. The result of this technique is excessive tracking error.
2. *Crosseye* (phase front distortion). Two widely spaced locations on the EA unit (such as the nose and tail or two wingtips in the case of an aircraft) are selected and interconnected transponders are installed. Each pair of these locations normal to the direction of the victim radar receives the victim radar pulse and triggers the transponder on the opposite side of the unit, which then transmits a copy of the victim radar pulse with a 180° phase shift.

The result is a reversal of the sign of the angular error measured at the victim radar. This causes the radar positioning mechanism to drive in the wrong direction. In the case of a TWS radar or an active tracking radar, this technique can result in errors in the positioning of tracking gates in bearing and elevation and can prevent the establishment of a smooth track, or can cause problems in acquisition gate, tracking gate, and turn-detection gate selection logic.

Continuous-wave Doppler and pulsed Doppler deception CW Doppler and pulsed Doppler radars are developed to track high-speed, low-flying aircraft in the presence of ground clutter. The echo return from these radars that enables the target to be tracked is the Doppler shift due to the target's velocity.

Deception of the CW Doppler requires that the repeater retransmit the received CW signal with a spurious Doppler shift, gradually increasing its magnitude to cause the velocity track to break. This will not only cause errors in the fire control solution, but because of the velocity gate walk-off, it can also result in loss of target tracking when the repeater is turned off.

Deception of the pulsed Doppler radar is much the same. The repeater introduces a similar spurious Doppler shift when it retransmits the received pulses.

Expendable Decoys

The next component of EA is the use of expendable decoys. As the name implies, expendable decoys are throwaway



FIGURE 6-10 *Super RBOC onboard USN combatant*

devices designed to imitate a real target. Virtually every platform in the Navy is capable of carrying and dispensing some type of expendable decoy; see figure 6-10. Expendables can be used under two tactical scenarios: as "force multipliers" masking real targets during offensive operations and as decoys, spoofing terminal threat emitters where the targeted platform is evading a weapon in the terminal phase of its guidance.

Chaff, flares, reflectors, and mobile or static offboard active transmitters are all examples of this type of EA device. Chaff rounds for aircraft are essentially oversized shotgun shells loaded with hundreds of small dipole antennas. Made of filaments of glass fiber coated with aluminum or zinc, these dipoles resonate when exposed to incident radiation of the appropriate wavelength. Essentially a small antenna, these filaments are cut lengths appropriate to the threat. Length (l) can easily be calculated using the ideal dipole antenna equation:

$$l = \frac{\lambda}{2} \qquad (6\text{-}4)$$

Typical dimensions for use against a 10 GHz radar would be 1.5 millimeters long, 0.02 millimeters wide, and .002 millimeters thick. Only 0.05 kilograms is needed to cause an echo equal in size to a large bomber.

Thousands of such dipoles are compressed into small packages. When injected into the air, the chaff packages burst open and the dipoles scatter to form a radar-reflective cloud called a chaff corridor. Chaff corridors move with the prevailing air currents and gradually disperse over time. Because they are slow-moving targets relative to the launch platform, they're prone to rejection by velocity-discriminating counter-countermeasures.

Surface ships carry larger chaff rounds to imitate their larger radar cross section. Generally, the dipoles in RBOC and superRBOC (super rapid blooming offboard countermeasures) rounds are longer, cut to the frequencies expected for antiship missiles.

Employment of chaff in and of itself is generally ineffective. Use of these countermeasures must be accompanied by platform maneuvers. Once countermeasures are deployed, the targeted platform must be maneuvered out of the field of view of the weapon's sensor while it is fixed on the decoy.

Reflectors are also used by both air and surface platforms to complicate air defense coordination or to confuse antiship missile seekers. The most common type used by surface platforms is known as the "Rubber Duck." The Duck is essentially an inflatable radar reflector that provides a ship-sized echo when illuminated by a radar. See figure 6-11.

The ADM-141 TALD (tactical air-launched decoy) is an airborne version of the Rubber Duck. Carried by the F/A-18 Hornet, TALDs are an active/passive decoy that can be programmed to glide to a specific profile and mimic an air wing strike.

Finally, there is a small category of active deception repeaters called echo or "blip enhancers." These repeaters strengthen the retransmitted pulse in order to make a small

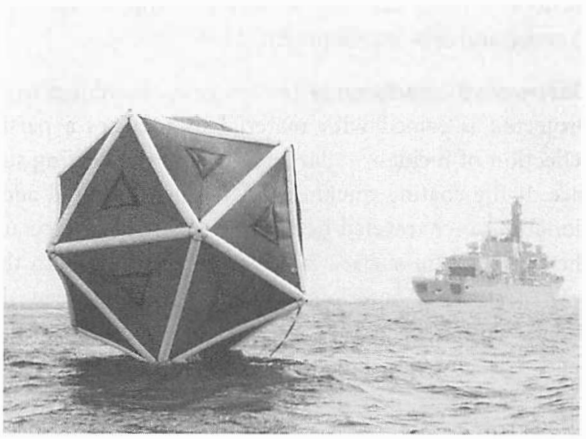


FIGURE 6-11 *Expendable floating decoys (courtesy Irvin-GQ Systems)*

radar target such as a destroyer appear as a large carrier. This may also be done mechanically by using properly designed reflectors that will make a small target look like a large one.

Target Masking and Modification (STEALTH)

Recent emphasis in the weapon field reaffirms the notion that "you can't hit what you can't see." A contact that remains undetected until its terminal phase carries the highest probability of successfully completing its mission. Toward that end, future weapon systems, and the platforms that carry them, will have to significantly reduce their distinguishable and detectable characteristics.

The classes of weapons so engineered are termed low observables and make use of the latest techniques and materials for reducing characteristic signatures. Because the primary focus of stealth technology is to avoid radar detection, it is included is here, but it applies to other segments of the EM spectrum and acoustic detection as well.

Techniques for modifying aircraft, ship, and submarine radar cross sections have existed since World War II. As discussed in chapter 3, radar cross section is a function of frequency, target size, shape, aspect, and composition. Reduction of radar cross section can be accomplished in many ways.

Alteration of platform shape The configuration of the platform must be modified according to the principles of geometrical optics such that large reflections are diverted to unimportant regions of space—that is, not back to the radar. Designers should avoid flat, cylindrical, parabolic, or conical surfaces normal to the direction of the radar illumination. These shapes tend to concentrate energy and provide a large radar return. Platform design should also include the use of doubly curved surfaces that result in low radar cross section. Unfortunately, in many cases these principles conflict with other important engineering requirements in aircraft and ship design, resulting in increased expense and slow development.

Destructive interference In this case, the object to be protected is coated with material that causes a partial reflection of incident radar energy from the coating surface. If the coating thickness is $\lambda/4$, then the total additional distance traveled between the coating surface and the object's skin surface is $\lambda/2$ or 180° in phase. In this way, destructive interference is achieved between the radar energy reflected from the coating surface and that reflected from the skin of the object. Thickness of the coating is not excessive at short wavelengths; however, it is impractical to install coatings thick enough to deal with low-frequency search radars.

Radar-absorbent material (RAM) Here, the object to be protected is given a coating of successive layers of magnetic composition material such as Ni-Mn-Zn sandwiched with dielectrics that convert 95% of incident RF energy to heat. This material can be made as thin as 1.75 centimeters, which is practical for aircraft use; however, the weight penalty of 24.9 kg/m^2 is excessive.

Another approach consists of a phenolic-fiberglass sandwich material. This structure again converts 95% of incident RF energy to heat by using a resistive material consisting of carbon and silver powder. This material is effective over the range of 2.5 to 13 GHz, which encompasses many fire control and weapon-guidance radars. The disadvantage of this approach is that while it is lightweight and relatively thin, it is not able to handle the high temperature and erosion processes at supersonic speeds. These methods, though promising, still cannot deal with some of the lower radar frequencies.

Antiradiation Weapons

The last method of EA is the use of antiradiation weapons. The only weapon of this type in U.S. inventory currently is the air-launched AGM-88 high-speed antiradiation missile (HARM). The HARM utilizes passive homing guidance to track the electromagnetic radiation emitted by enemy radar. The missile is preprogrammed with the parametric data of the target prior to engagement.

Once the radar is detected, the missile is launched at ranges out to approximately 80 nautical miles and homes on radiation at the specific parameters programmed into its guidance section. Its fragmentation warhead contains more than 10,000 tungsten steel cubes intended to shred the antenna and wave guides of the target. Often, the threat of a HARM launch will cause an enemy to shut down acquisition, tracking, and guidance radars in order to avoid damage.

Electronic Protection

Electronic protection is the art of reducing the effectiveness of an EC threat with the objective of making the cost of effective EC prohibitive for the enemy. As in EA, EP includes both radar design and operator training. EP designers must understand the various forms of EA that a radar is likely to encounter; hence, intelligence about the EA threat is very important. Therefore, the designer must provide a variety of options to be used against the expected threats.

The operator must also be trained both to recognize the various countermeasures that might be used and to select the appropriate combination of options to counter each of them.

TABLE 6-1 *Summary of EA techniques by class and type*

Class	Technique	Description
1. **Active radiators**—radiate active signals to interfere with the radar	Denial Jamming	
	• Barrage	Covers wide bandwidth
	• Spot	Covers narrow bandwidth
	• Sweep	Sweeps signal over wide bandwidth
	Deception Jamming	
	• Range	Changes range of target
	• Angle	Changes angular position of target
	• Frequency	Changes velocity of target
2. **Medium modifiers**—change the electrical properties of the medium between the aircraft/ship and the radar	Chaff	Provides numerous returns
	Expendable decoys	Provides large returns
3. **Reflectivity modifiers**—change the reflective properties of the aircraft or ship itself	RAM	Absorbs radar energy
	Vehicle design	Scatters radar energy
	Echo enhancers	Provides more return to radar
4. **Physical destruction**—uses a weapon to damage the transmitting radar itself (hard kill)	Antiradiation missiles	Destroy antenna-transmitting equipment

The most effective measure to combat EA is an up-to-date piece of equipment operated by a well-trained operator.

Radar design for EP can be broken down into three areas: radar parameter management, signal-processing techniques, and radar design philosophy.

Electronic protection also includes the passive detection and home-on jam techniques where the "radar" does not transmit, but uses the EA energy emitted by its victim to determine location, often by triangulation from two or more separate locations.

Radar Parameter Management

The basic radar parameters that are managed are those characteristics that influence the radar's performance. These are power, frequency, PRF, pulse length, antenna gain, antenna polarization, antenna scan, and antenna sidelobe characteristics. These values, and the means with which they can be manipulated, are established in the design phase.

Power For ground or surface radar, power is often considered the fundamental EP parameter. With this view, EA becomes a power battle, with the outcome going to the stronger, more powerful opponent. Airborne jamming equipment is limited in size and weight and therefore has a power limitation. Thus, the power advantage lies with the ground or surface radar. In the case of one surface unit versus another, both operate under the same constraint, and the outcome is not obvious.

Frequency Frequency agility is a significant EP design feature. Using components such as frequency synthesizers similar to those employed in radio scanners, instead of conventional crystal-controlled oscillators, some radars are able to change frequency within one pulse repetition time (PRT). This makes deception and jamming very difficult. The radar can be designed to change frequency automatically or manually.

A second way of using frequency as an EP technique is Doppler radar, including radars designed for MTI signal processing. The actual EP advantage is gained from signal processing in the receiver, but the intention to use the Doppler frequency shift must be reflected in the transmitter design. For example, in a pulse-Doppler radar, the transmitter must often be designed to radiate a very stable frequency. In a pulse compression radar, the transmitter must radiate a pulse with an FM slide, sometimes called a chirp pulse due to the change in "tone" as the pulse is transmitted.

Pulse repetition frequency (PRF) In general, high PRF radars are more resistant to EA because the average power is greater. Changing the PRF in a random fashion is an

effective counter to deception because EA often depends on predictability of the radar. However, because PRF is related to the basic timing of the radar, this technique results in additional complexity and expense. Random PRF has been employed as a very effective EP feature in some radars for many years and has the additional benefit of eliminating radar blind speeds as discussed in chapter 4.

Pulse length An increase in pulse length will increase average power and thus increase detection probability. The trade-off is increased minimum range and degradation of the radar's range resolution capability. This problem is compensated for by including a pulse compression capability; however, due to receiver blanking during the transmit cycle, the minimum range will stay relatively long.

Some modern radars compensate for these difficulties by employing pulse compression and varying their pulse width depending on mode of operation and expected target range.

Antenna design Antenna design, as reflected in low sidelobe levels, is an EP design technique, because it prevents a jammer or deceiver from affecting the radar at many bearings. Low sidelobe levels also make the job of antiradiation missiles more difficult, since there is less chance of the missile homing in on the radar unless the radar is pointing at the missile. Sidelobe patterns can be compensated for by employing techniques such as sidelobe cancellation and sidelobe suppression.

Scan pattern The radar scan pattern can influence EP capability because it influences the amount of energy directed toward the radar target. An active tracking phased-array radar is very EA resistant because of its ability to rapidly scan its radar beams in a random fashion rather than in the usual circular or sector scan pattern of conventional radars. This irregular beam positioning would give the opposing EA system little or no warning and make it impossible to predict where and when to transmit false signals.

In systems where scanning is performed in the receiver rather than in the transmitted beam, such as those mentioned in the section on angle deception, EA has no direct access to the radar scan pattern and thus has difficulty using that information to interfere with the radar system operation.

Signal-Processing Techniques

These are usually functions that are incorporated into the radar receiver. Although certain signal-processing techniques may place constraints on the transmitter, many of them have been added to the receiver after the radar has been built. These techniques are called EP or antijamming (AJ) fixes, since they were initially developed as retrofits to improve existing equipment. Radars now tend toward a more sophisticated design concept in which the antijamming devices are included in the basic radar system.

Doppler radars including radars with MTI signal processors, although not designed specifically for EP purposes, are quite EA resistant. Since Doppler radars, both pulse and CW, operate on the frequency shift caused by a moving target, they automatically filter out returns from nonmoving targets and consequently eliminate many unwanted signals, such as those from chaff. They will even discriminate between returns from objects of different velocities such as an aircraft in a chaff cloud. This technique can also make deception more difficult, since the deceiver must imitate the proper frequency shift.

In radars that have automatic threshold detection where the target is displayed when receiver output crosses a preset threshold, the presence of a jamming signal can increase the rate of false targets to an intolerable extent. If the radar output data are processed in an automatic device such as a computer, the device might be overloaded by the added false targets caused by jamming.

Thus, it is important that the receiver present a constant false-alarm rate. Receivers designed to accomplish this are called CFAR (constant-false-alarm rate) receivers. Their disadvantage lies in the likelihood that some weak targets will remain below the threshold and be lost.

If an operator were monitoring the radar output, the effect of the additional false alarms could be reduced by having the operator turn down the gain of the receiver during the presence of jamming, or ignore those sectors containing EA.

In an automatic threshold detector, the same effect may be obtained by using the average noise level to provide an automatic gain control, much as an operator would by adjusting a manual gain control. Because the automatic CFAR circuits react faster, they are superior to an operator in keeping the false-alarm rate constant, especially when the radar is subject to noise jamming from only a few bearing sectors.

A CFAR receiver, no matter whether it is an automatic device or an operator controlling the receiver gain, maintains the false-alarm rate constant by reducing the probability of detection. However, when the threshold level is raised to maintain a constant false-alarm rate, marginal echo signals that might normally be detected do not cross the higher threshold and are lost.

Therefore, CFAR does not give immunity to jamming; it merely makes operation in the presence of jamming more convenient by making the receiver less sensitive. If the

jamming were severe enough, the CFAR, for all intents and purposes, could produce the same effect as the operator's turning off the receiver.

Other EP signal-processing techniques, such as those discussed in chapter 5 concerning track-while-scan radars, employ track history to reject false returns. The addition of computerized processing of radar video can perform this function as well as echoes returned in several successive radar transmissions as a means to counter the deception jammer.

Radar Design Philosophy

A general rule of thumb for EP radar design is to incorporate unpredictable operating parameters. The more orderly a radar is in its operation, the easier it is to predict what the radar is going to do or how it is going to operate; consequently, the job of applying an EA technique effectively becomes simpler.

EA becomes more difficult, however, if characteristics of the victim radar are constantly changing. The parameter that may most easily be varied to confuse the EA operator is the frequency. The capability for operator variation of pulse length, PRF, modulation, and antenna characteristics is commonly built into radars to make EA more difficult.

The most common way to introduce unpredictability into radar design is through frequency diversity. Early radars were all designed to operate in a few specific frequency bands, where narrowband jamming would render them all ineffective. New radar systems are designed so that each different radar type operates in a different frequency band.

The use of a much greater portion of the spectrum, from VHF to SHF (A- to J-band), forces EA operators to cover this total spectrum to be effective. This usually results in their having to put less EA power against a single radar because of the limit in total power-out capability.

Another aspect of EP design philosophy is the relationship between automatic equipment and the human operator. The trained radar operator fulfills a useful and necessary role in a countermeasure environment and cannot be completely replaced by automatic detection and data processors.

An automatic processor can be designed to operate against only those interfering or jamming signals known beforehand; that is, any capability against such signals must be programmed into the equipment beforehand. New jamming situations not designed into the data processor might not be readily handled.

On the other hand, a human being has the ability to adapt to new and varied situations and is more likely to be able to cope with, and properly interpret, a strange new form of interference than can a machine. Therefore, a skilled operator is the most important counter-countermeasure for maintaining radar operation in the presence of deliberate and clever countermeasures.

Anti-ES Countermeasures

Radar, fire control, and weapon sensors tend to dominate the electronic counter-countermeasures area. There are, however, certain tactical approaches that should be considered when trying to counter an opponent's EC capability, such as RF emission control (EMCON) and evasion of the opponent's ES or active sensors. Application of these tactics must be tailored to the mission objective since each requires giving up information and possibly some operational flexibility.

EMCON is a situation of degree as opposed to no emissions at all. The price of complete EMCON in terms of absence of communications and lack of tactical data may considerably inhibit mission completion. Limited use of surveillance sensors, especially those that are common enough not to identify a ship or ship class, can provide adequate tactical information without divulging force size or capability.

Use of sensors in other parts of the electromagnetic spectrum such as passive IR or even passive acoustics is an alternative to active RF sensing. A combination of the long-range capability of passive RF sensing with the high accuracy of passive IR can provide early warning, identification, and precise angular location data on platforms of interest.

Line-of-site communications (i.e., VHF, UHF) are far less prone to interception than HF communications, which follow the earth's curvature. If directional communication capability is available, it may be used to further increase intercept resistance. Even though platform location may be known to opponents, intentions may be concealed by restricting fire control sensor emissions and minimizing communication traffic until such emissions are absolutely necessary.

EMCON may also be exercised by limiting the time and regularity of key active sensor operation. Regular repetition of emissions is of great benefit to the ES identification process and deceptive countermeasures. Toward this end, LPI, or low probability of intercept, communications will become the rule rather than the exception.

Evasion, although limited by mission characteristics, is still a mission planning consideration. Aircraft may use terrain or the radar horizon as a shield. Nondirect flight routes by aircraft may well be worth the flight time, fuel, and reduced weapon load-out in terms of possible aircraft losses due to an alerted opponent. Radar horizon protection is also available to surface combatants.

The Effectiveness of EA/EP

The interaction between friendly and hostile electronic combat is analogous to a ladder of sequential increases in technological capability that alternate between two opponents. Fielding of an electronic system by one side results in the development of a countering system by an opponent, which in turn causes the original side to develop a counter-countermeasure. The process continues to escalate through a series of countermeasures and counter-countermeasures, a "ladder." This process can be diagrammed as shown in figure 6-12.[iv]

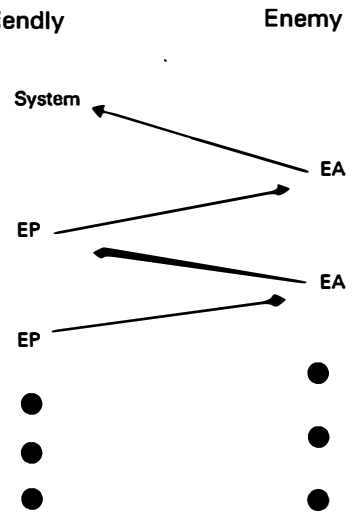


FIGURE 6-12
EA ladder

This analysis shows that unequivocal superiority can never be achieved through EA. Unfortunately, this analysis tends to promote a two-valued evaluation of EA: EA either works or it doesn't, depending on what rung of the ladder the technique is currently on. However, a more realistic evaluation of this chain yields the following conclusions:

1. The real effectiveness of EA lies somewhere on a spectrum ranging from completely effective to completely ineffective. The actual evaluation depends both on the position on the ladder and on many other factors, some relating to training and morale and others relating to the uncertainties of combat.
2. EA techniques have only a finite time of superiority. Eventually an adversary will develop a countertechnique and the superiority will pass. Thus, one cannot expect a certain EA technique to give indefinite superiority.
3. The real advantage of EA, then, is that it gives relative superiority while an adversary is developing and deploying a countermeasure. Unless an adversary is very perceptive, this relative advantage can be measured by the time delay between the operational employment of two successive steps on the ladder. During this time, the succeeding idea is being developed and the necessary equipment being built. Therefore, good security is necessary to preserve the relative advantage.
4. Technological superiority and constant development in the EC area are required to be consistently able to counter enemy advances.
5. Techniques should not be discarded because they have simple counters; there will always be benefits from the inevitable delay between the time the enemy is certain that the technique is being used and the time needed to make the countermeasure operational.

Summary

It is imperative that professional military officers have a basic understanding of this warfare area. This chapter was designed to create an awareness and to assist the reader in developing an appreciation for some of the capabilities, limitations, and applications of electronic combat. It is included with sensor systems because electronic combat is directly interrelated with sensors and in one sense, is itself an extremely valuable sensor, frequently providing the commander with the initial enemy detection. The importance of electronic combat in today's highly complex and heavily electronically dependent order of battle cannot be overemphasized.

---

[i] *FAS*
[ii] *FAS*
[iii] *FAS* Superheterodyne
[iv] *FAS*

7 Principles of Electro-Optics Sensing

Introduction

The first optical sensor of naval importance was undoubtedly the human eye. Since the revolutionary improvement of the telescope in 1609 by the Italian physicist and astronomer Galileo, optics has played a major role in naval strategy and tactics. In fact, when Galileo first presented his original eight-power telescope, its value for naval and maritime operations was quickly recognized as many Venetian senators climbed the highest bell towers in Venice and observed ships far out at sea.[i]

For centuries, the limiting factor for tactical application of optical systems was the requirement for an illumination source. Although military tacticians throughout history have always seen the advantages of being able to maneuver effectively under cover of darkness, maneuvering large forces at night carried such risks that it was rarely attempted.

Early efforts to eliminate the hazards of darkness included the use of torches, flares, and rockets. It is easy to understand the appeal that electro-optical technologies have for the military. Since infrared radiation (IR) cannot be detected by the human eye, it offers the possibility of detecting targets by their self-emission of electromagnetic radiation, thereby "seeing" in environments where the eye is inefficient.[ii]

Prior to World War II, naval optical systems were important in fire control, weapon delivery, surveillance, and reconnaissance. Many of these functions were then taken over by electromagnetic sensing at radio frequencies, and although the use of optics was eclipsed, it was never abandoned. In fact, the militaries of the United States, Great Britain, and Germany began experimenting with infrared equipment soon after 1900.

During World War I, formal research work was initiated in the realm of night vision, but such efforts were devoted primarily to searchlight illumination of scenery. However, it was during this period that experimental blinker signaling, and search equipment were developed and proved successful. Although neither side put any infrared equipment into production before the end of the war, the experience gained with the experimental equipment clearly demonstrated the military potential of infrared technology.[iii]

The development of the infrared image converter tube on the eve of World War II made it possible for Germany to be the first country to deploy infrared equipment in the field. Early in the war, the Germans concluded that the Allies were using infrared equipment to detect Nazi U-boats and aircraft. Because of this erroneous conclusion, the Germans concentrated much of their infrared research and development on the means of countering the Allies' employment of such devices.[iv]

In the United States during World War II, the Office of Scientific Research and Development (OSRD) oversaw the electro-optical effort. It was television research by RCA (the Radio Corporation of America) in the 1930s that led to the development of an image tube that could be used to convert infrared images into visible displays.

This development of a night vision device culminated in 1944 with the introduction of an operational small-arms-mounted infrared unit called the Sniperscope. This first IR Sniperscope used near-infrared cathodes coupled to visible phosphors to provide a near-infrared image; see figure 7-1 and 7-2. First used during the Okinawa Campaign, the Sniperscope's night vision capability proved particularly effective in combating Japanese night infiltration tactics.

Classified "secret" at the time, information on the infrared equipment and its combat effectiveness was not disclosed until after the war. The particular effectiveness



FIGURE 7-1 *World War II infrared M1 "Sniperscope" mounted on a .30 caliber carbine TM3*



FIGURE 7-2 *Marines of the 1st Regiment, 1st Marine Division, testing the improved M3 "Sniperscope"*

of this night vision capability is supported by combat reports citing approaching troops being thoroughly decimated while attempting to pick their way through the defensive lines.[v]

The original IR Sniperscopes did have significant problems that caused them to be used mainly in perimeter defense: the effective range was typically less than 100 yards; the rifle-mounted scope required cumbersome batteries; and since it was coupled with an active IR searchlight, it could also be detected by any enemy soldiers equipped with similar devices. However, military leaders immediately saw many uses for this technology beyond simply sniping at the enemy under cover of darkness. The next challenge in night vision technology would be the development of passive systems that did not require an IR searchlight that might give away a soldier's position to the enemy.[vi]

Electro-optical sensing systems are gaining prominence in naval warfare applications today. Developing electro-optical (EO) technology includes devices and techniques for passive sensing and imaging of optical and IR energy from targets, and extends the techniques and capabilities of radar into the visible and IR frequencies. Additionally, laser weapons that are able to deliver high quantities of energy to targets are now feasible. It will be seen that the advantages derived from the use of optical and infrared frequencies, which include precision imaging, tracking, location, and ranging, often offset the disadvantages related to limited ranges due to atmospheric absorption and scattering.

This discussion of EO systems is structured to systematically examine the process of how EO energy is created, transmitted, and then ultimately received by a detection sensor. The "system" comprises the radiation sources, targets and backgrounds, the atmosphere, and the sensor; see figure 7-3. The sensor includes optics, scanner (if required), detectors, electronics, display, and human perception. If an automated target recognition processor is used, the display and human perception are not required.[vii]

The Electro-Optical Spectrum

The portion of the electromagnetic (EM) spectrum that lies between X-rays and microwaves is referred to as the electro-optical spectrum. This small band encompasses the ultraviolet region, visible region, and infrared region. The visible band is generally centered at a frequency of approximately 5.4×10^{14} Hz, which is in the "green-yellow" response of human optical vision.

Frequencies ten times lower than this are in the invisible "infrared," usually thought of as heat radiation, and frequencies ten times higher than this are in the invisible "ultraviolet." The expanded scale of figure 7-4 shows a portion of the EM spectrum significant to this discussion.

Because frequencies in this portion of the spectrum are in the millions of megahertz (MHz), it is customary to refer to wavelength rather than frequency when describing the specification of these waves. The unit most commonly used to describe wavelength of visible and IR energy is the micrometer (μm) or "micron," which is 10^{-6} meter long. Note that since there are no clear-cut boundaries to the various portions of the electromagnetic spectrum, the regions may slightly vary from publication to publication.

However, it is generally accepted that visible light extends from approximately 0.4μm to 0.7μm, whereas IR extends from 0.7μm to 1,000μm. Visible light can be further broken down into bands according to color, and IR

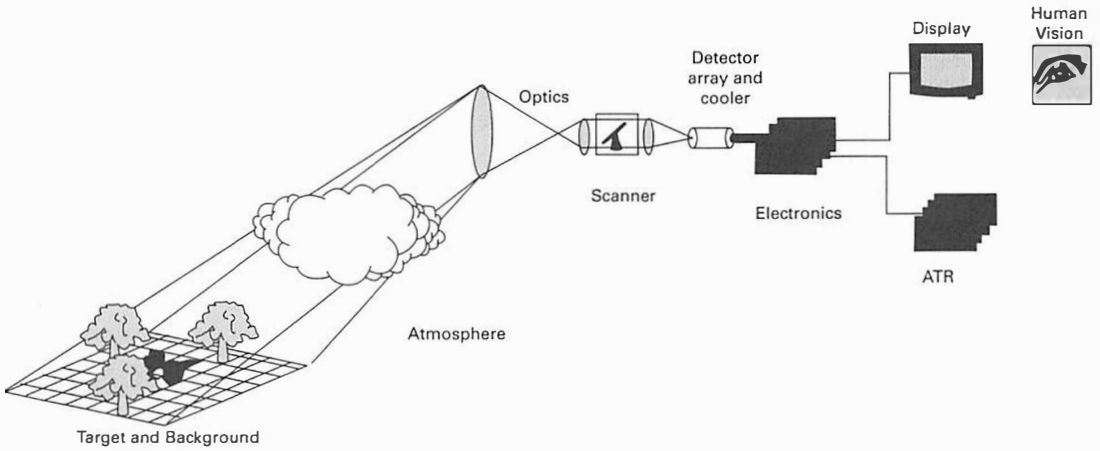


FIGURE 7-3 *A generic electro-optic or thermal imaging "system"*

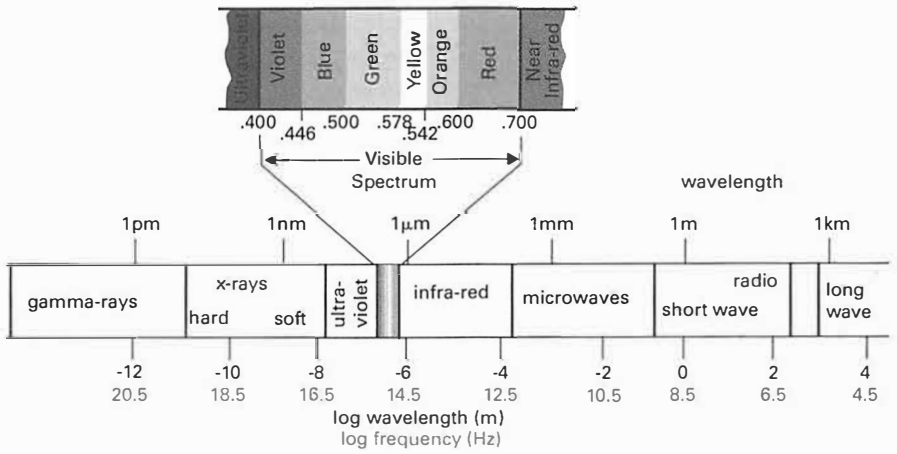


FIGURE 7-4 *Electro-optic spectrum*

into separate bands called near or "shortwave" IR (NIR or SWIR), middle or "midwave" IR (MIR or MWIR), far or "longwave" IR (FIR or LWIR), and extreme IR (XIR). See table 7-1 for a specific break down of wavelengths for each of the above bonds.

TABLE 7-1 *Wavelengths of visible and infrared radiation*

Visible Light ($\lambda = 0.4 - 0.7\mu m$)	
Infrared Regions ($\lambda = 0.7 - 1{,}000\mu m$)	
• **NIR or SWIR**	0.7–3 μm
• **MIR or MWIR**	3–6 μm
• **FIR or LWIR**	6–15 μm
• **XIR**	15–1000 μm

The logic of these four choices becomes apparent when compared with the useful atmospheric transmission "windows" or readily available detector response bandwidths. For example, the first three subdivisions include regions in which the earth's atmosphere is relatively transparent; in the last subdivision, the atmosphere is essentially opaque for paths more than a few meters long.[viii]

For purposes of this text, visible imaging systems or image intensification systems are considered to be those that detect energy in the 0.4–2μm wavelength band and infrared imaging systems to be those that detect energy between 3μm and 14μm. This creates two categories of sensor systems, those that primarily detect and amplify visible energy reflected off scenes—that is, night vision devices—and those that primarily detect energy radiated or emitted from a target—that is, infrared or thermal imaging.

Radiometry vs. Photometry

The principles inherent to the measurement of light evolved many years before it was known that light was not a separate entity but merely a narrow region of the EM spectrum. Radiometry is the measurement of all "electro-optical radiation," which is EM radiation within the frequency range between 3×10^{11} and 3×10^{16} Hz. This range corresponds to wavelengths between 0.01μm, and 1,000μm and includes the regions commonly called the ultraviolet, the visible, and the infrared.

Photometry is the measurement of light, which is defined as electromagnetic radiation that is detectable by the human eye. It is thus restricted to the wavelength range from about 0.36μm to 0.83μm. The only real difference between radiometry and photometry is that radiometry includes the entire optical radiation spectrum, whereas photometry is limited to the visible spectrum as defined by the spectral response of the eye.[ix]

Visual photometry uses the eye as a comparison detector, and physical photometry uses either optical radiation detectors constructed to mimic the spectral response of the eye, or spectroradiometry coupled with appropriate calculations to do the eye-response weighting. Typical photometric units include lumens, lux, and candelas.[x]

Table 7-2 displays common quantities used in photometry. The military typically uses photometric units when specifying sources of illumination such as the moon, stars, or high-intensity illumination flares for night vision devices (NVD), also called image intensifiers.

TABLE 7-2 *Photometric units*
Note: E. unit of sr stands for steradian

Term	Symbol	Units
Luminous flux	Φ	Lumens, (L)
Luminous intensity	*I*	(L/sr)
Luminous exitance	*M*	(L/m²)
Illuminance	*E*	(L/m²) or lux
Luminance	*L*	(L/m² · sr)

Plane Geometry vs. Solid Geometry

Radian

The radian is the Standard International System (SI) unit of plane angular measure. There are 2π, or approximately 6.28, radians in a complete circle. Thus, 1 radian is about 57.3 angular degrees. The term *radian* arises from the fact that the length of a circular arc, corresponding to an angle of 1 radian, is equal to the radius of the arc. This is shown in figure 7-5. Point *P* represents the center of the circle. The angle θ, representing 1 radian, is such that the length of the subtended circular arc is equal to the radius, *r*, of the circle.

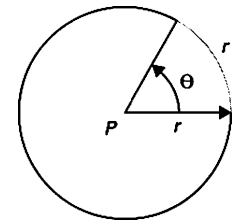


FIGURE 7-5 *Radian defined*

Steradian

The steradian is the SI unit of solid angular measure, symbolized by "sr." There are 4π, or approximately 12.57, steradians in a complete sphere. Just as a radian is a measure of the angle subtended by a one-dimensional line about the origin in two-dimensional (plane) space, steradians are a measure of the angular area subtended by a two-dimensional surface about the origin in three-dimensional space.

A steradian is conical in shape, as shown in figure 7-6. The solid angle (conical angle) Ω, representing 1 steradian, is such that the area *A* of the subtended portion of the sphere is equal to r^2, where *r* is the radius of the sphere and point *P* represents the center of the sphere.

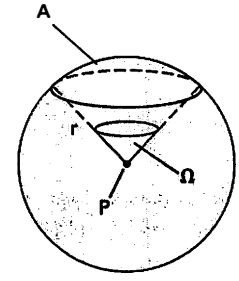


FIGURE 7-6 *Steradian defined*

The numerical value of a solid angle is a given area on the surface of the sphere divided by the square of the sphere radius; thus, a solid angle is given by:

$$\Omega = \frac{A}{r^2} \qquad (7\text{-}1)$$

where:

Ω = solid angle (measured in steradians)
A = curved surface area of the sphere
r = radius of the sphere

Consider again the sphere in figure 7-6. The maximum number of steradians on a spherical surface is $4\pi r^2/r^2 = 4\pi$ sr. If the radius is much larger than any dimension of the surface area A, then the area on the sphere can be assumed to be flat. Given any flat shape imaginable and the distance to the shape, a steradian can be calculated for the corresponding solid angle using equation 7-1.[xi] Note that steradians are dimensionless.

Steradians and solid angles are parameters that must be understood because many radiometric units are given with respect to solid angles.

Radiometric Quantities

Quantitative characterization of radiant sources requires a precise system of units. Historically, many different radiometric terms have been introduced and used; however, this text will use the approved (SI) units. These terms and their units are summarized in Table 7-3.

The radiometric units listed in Table 7-3 are based on the Joule as the fundamental quantity of energy, and are thus termed energy derived. An analogous set of quantities, based on number of photons, is termed photon-derived or photometric. The photon-derived quantities are typically written with a subscript q. The units that are energy derived are written with a subscript e (electromagnetic) only if it is necessary to distinguish between photometric units. Typically, radiometric quantities are used for sensing energy in the IR spectrum, and photometric quantities are used when dealing with sensors designed for the visible spectrum.

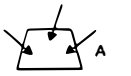
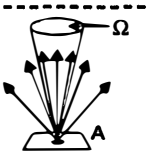
Radiant Energy—Q

Radiant energy is energy transported in the form of electromagnetic waves from a source. The term is used to describe the entire amount of energy radiated from a source in a given time interval. Radiant energy is usually expressed in Joules (J).

Radiant Flux (Power)—Φ

Radiant flux, also called radiant power, is the rate of flow of radiant electromagnetic energy. Radiant power is usually expressed in watts (w). In many cases the modifier is dropped and "power" is used to mean "radiant power." Radiant flux (Φ) is energy per unit time that is radiated from a source. Therefore, if radiant flux is integrated over

TABLE 7-3 *Radiometric units*

Term	Symbol	Units	Defining equation	Meaning	Example Application
Radiant energy	Q	Joule, J (w · s)			
Radiant flux (power)	Φ	watts = J/s or photons/sec	$\Phi = \frac{dQ}{dt}$		Power on a detector
Intensity	I	w/sr	$I = \frac{d\Phi}{d\Omega} \Rightarrow \frac{\Phi}{\Omega}$		Target signature for non imaging seeker
Radiant exitance	M	w/m² or w/cm²	$M = \frac{d\Phi}{dA} \Rightarrow \frac{\Phi_{180°}}{A}$		Radiative cooling
Irradiance (incidence)	E	w/m² or w/cm²	$E = \frac{d\Phi}{dA} \Rightarrow \frac{\Phi}{A}$		Most IR receivers; heat transfer analysis
Radiance	L	w/(m² · sr) or w/(cm² · sr)	$L = \frac{dI}{dA} \Rightarrow \frac{\Phi}{A\Omega}$		Target signature for imaging sensors

time, the total radiant energy output (Q) from the source is obtained.

Radiant Exitance—M

Exitance and irradiance have the same units, but different interpretations. Exitance is the amount of power per unit area expressed in watts per square meter (w/m^2), or square centimeter (w/cm^2), that leaves a surface and is radiated into a hemisphere. Exitance thus characterizes a self-emitting source that is producing energy and is written in differential form as $M = \frac{\partial \Phi}{\partial A}$ and can be approximated for small areas as

$$M = \frac{\Phi}{A} \qquad (7\text{-}2)$$

EXAMPLE 7-1: If a 1 mm^2 radiating surface has an exitance of 3 w/cm^2, what is the total radiated power?[xii]

SOLUTION:

$$\Phi = A_s \times M$$
$$= (0.1\,\text{cm})^2 \times 3\ \text{w/cm}^2$$
$$= 30\ \text{mw}$$

where:

Φ = radiant flux or power emitted (w)
A_s = surface area (cm^2)
M = radiant exitance (w/cm^2)

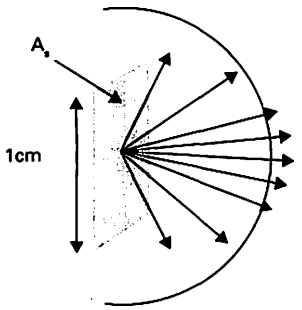


FIGURE 7-7 *Exitance of a source*

This power is radiated into the entire hemisphere in front of the source, but since it is not a point source, it is not radiated uniformly in angle; see figure 7-7.

Irradiance—E

Irradiance is the radiant power per unit area incident upon a surface, regardless of direction of arrival, and is usually expressed in watts per square meter (w/m^2) or square centimeter (w/cm^2). Irradiance characterizes a passive receiving surface, in contrast to exitance, which characterizes a self-emitting source producing energy.

EXAMPLE 7-2: Assume that the irradiance falling on a 1 mm^2 surface is 4 w/cm^2 and assume it is uniform over the surface; see figure 7-8. How much power is received by a 1 mm × 1 mm portion of the receiving surface?

SOLUTION:

$$\Phi = A_s \times E$$
$$= (0.1\,\text{cm})^2 \times 4\ \text{w/cm}^2$$
$$= 40\ \text{mw}$$

where:

Φ = flux or power incident on surface (w)
A_s = surface area (cm^2)
E = irradiance (w/cm^2)

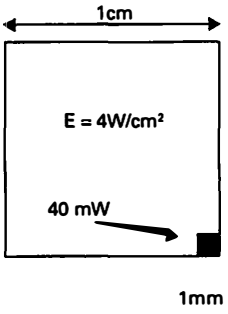


FIGURE 7-8 *Irradiance of a receiving surface*

Point Source vs. Extended Source

Before discussing the radiometric measurements of radiant intensity and radiance, it is necessary to differentiate between point sources and extended sources. A true point source is not physically realizable, but it can be closely approximated by a small source at a great distance, that is, a star. A point sources sometimes called an isotropic radiator, which radiates equally in all directions in three-dimensional space. This is in contrast to an extended source, which is an object that produces an image at the detector optics larger than the detector itself.

What is important is not the physical size of the source but the angle that subtends the detector. Therefore, the same source can at different times be classed as either point or extended. For example, a tailpipe of a jet aircraft at a distance of 10 kilometer is effectively a point source, whereas at a distance of 50 meter, it represents an extended source.

If the radiometer uses optics, the basic criterion is the relationship between the size of the detector and the size

of the source image. If the image is smaller than the detector, the source is considered to be a point source. If the image is larger than the detector, the source is considered to be an extended source. In an effort to quantify the power emitted by a source, a distinction is made between types of sources, and separate terms are used to describe the radiation from each.[xiii]

Radiant Intensity—I

Radiant intensity is an expression of the brilliance of a source of EM energy over the entire electromagnetic spectrum. This quantity is measured in terms of the flux or power emitted per unit steradian from an isotropic radiator. Intensity is the only quantity that can be used to specify the radiation from a true point source. Since a point source has negligible area compared to the square of the viewing distance, any quantities with the source area in the denominator cannot be used.

Intensity can also be used to specify the watts per unit solid angle radiated from an extended source; however, this intensity will vary with the viewing angle of the detector, as shown in figure 7-7.[xiv] The standard expression of radiant intensity is watts per steradian (w/sr) and is found by measuring the radiant flux and dividing it by the solid angle subtended by the detector, at the source.

EXAMPLE 7-3: Consider a point source with a uniform intensity of 3 w/sr. What is the total radiated power?

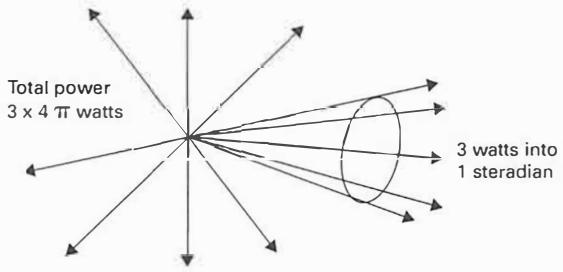


FIGURE 7-9 *Intensity of a point source*

SOLUTION: If the total solid angle into which a point source radiates equals 4π sr, then:

$$\Phi = I \times \Omega$$
$$= 3 \text{ w/sr} \times 4\pi \text{ sr}$$
$$= 12\pi \text{ w}$$

where:

Φ = radiant flux or power (w)
I = intensity of the point source (w/sr)
Ω = the solid angle (sr)

Radiance—L

If the source is an extended one—that is, a source that has appreciable area compared to the square of the viewing distance such as the sky or ground—it is not possible to define the solid angle subtended by the detector at the source. This difficulty is resolved by characterizing an extended source by its radiance. Radiance is the amount of power radiated per unit projected source area per unit solid angle from the given direction.

Radiance is usually expressed in watts per steradian per square meter (w/sr · m^2) or square centimeter (w/sr · cm^2). In order to measure radiance, it is necessary to limit the measurement to a small area of the extended source. In effect, it is the measurement of the radiant intensity of a small, but known, portion of the extended source.

Since flux is measured from a distance, the exact area involved in the measurement, and its orientation, is difficult to determine. However, if the source area is defined as the area projected on a plane located at the source and normal to the direction of measurement (viewing direction), the result is a quantity that can be measured with ease, even from a distance.[xv] The projected area is equal to the actual surface area times the cosine of the angle between the normal to the surface and the viewing direction.

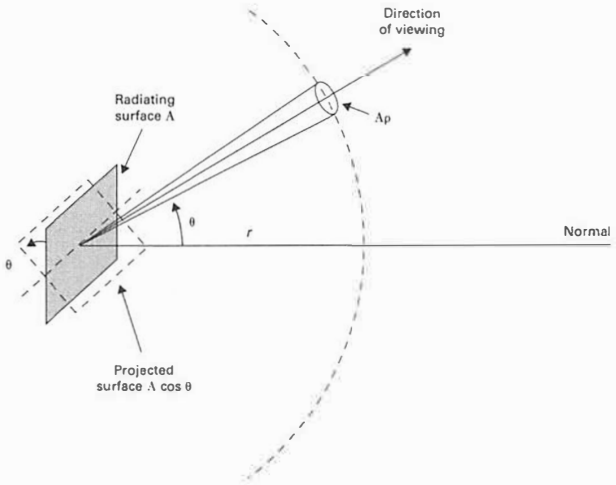


FIGURE 7-10 *Radiant intensity collected along a direction* θ

The importance of the radiance is suggested in the following considerations. Suppose a plane radiator radiates uniformly in all directions. The radiant intensity can be measured by the fixed aperture A_p at a fixed solid angle θ at some distance r from the radiating surface, shown in Figure 7-10.

The aperture, A_p, may be considered the input aperture of a detecting instrument measuring all the flux that enters. When viewed at $\theta = 0$, normal to the radiating surface, a certain maximum intensity $I_{(0)}$ is observed. As the aperture is moved along the circle of radius r, thereby increasing the angle θ, the cross section of radiation presented by the surface decreases in such a way that:

$$I_{(\theta)} = I_{(0)} \cos \theta \qquad (7\text{-}3)$$

where:

I = intensity measured by the aperture (w/sr)
θ = angle measured from the normal to the radiating surface (rad)
$I_{(0)}$ = intensity measured when $\theta = 0$

For a perfectly diffuse reflector (one that scatters energy in all directions equally), the reflected flux per unit solid angle is proportional to the cosine of the angle between the direction of interest; this relation is called Lambert's cosine law. (Equation 7-3 is Lambert's cosine law.) It simply states that the reflected or transmitted intensity in any direction from an element of a perfectly diffusing surface varies as the cosine of the angle between that direction and the normal vector of the surface varies.

Although Lambert's cosine law is an idealized concept, many sources follow it quite closely. The blackbody, which will be discussed in detail later, follows it exactly. Most materials that are electrical insulators follow it until the viewing angle exceeds 60° from the normal; materials that are electrical conductors show somewhat greater departure, but until the viewing angle exceeds 50°, the cosine law is useful in engineering calculations.[xvi]

Therefore, if the radiance measured at each angle θ is found to be constant, the surface is described as Lambertian. This can occur because the intensity must be divided by the projected area, $A \cos \theta$, such that the cosine dependence cancels:

$$L = \frac{I_{(0)} \cos\theta}{A \cos\theta} = \frac{I_{(0)}}{A} = \text{constant} \qquad (7\text{-}4)$$

where:

L = radiance of the surface (w/(cm^2 · sr))
$I_{(0)}$ = intensity of radiating surface (w/sr)
A = surface area (cm^2)
θ = angle measured from the normal to radiating surface (rad)

Lambertian Surface

If a beam of incident radiation is reflected strongly into a single direction such as with a mirror, the reflection is called specular and the surface is described as glossy. With this condition it is relatively easy to collect and measure all of the reflected flux. If the radiation is reflected into all directions with a uniform distribution, the reflection is called diffuse and the optical surface is sometimes described as lusterless. Since the reflected flux is spread over a wide solid angle, it may be difficult to ensure that all radiation is collected and measured; see figure 7-11. Most real surfaces fall somewhere between the two extremes.

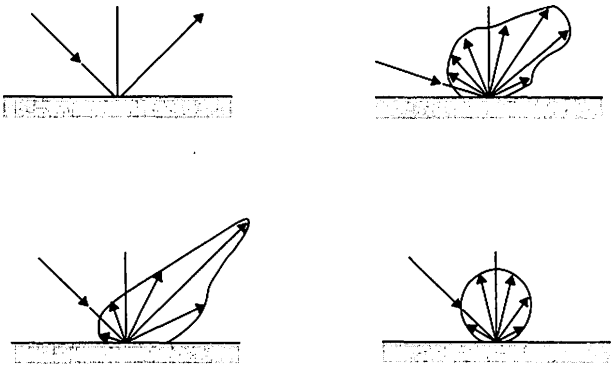


FIGURE 7-11 *Variations in directional reflectivity*

The ideal "diffuse" surface is one in which the angular intensity profile varies as the cosine of the angle from the surface normal, again Lambert's cosine law. These are called Lambertian surfaces and although many materials appear to be Lambertian, few surfaces actually are. In the visible spectrum such diffuse surfaces are called matte or flat. The Lambertian surface approximation applies to a majority of natural and human-made surfaces including most rough surfaces such as paints, cloths, stones, and wood that is not shiny. It is important to note that roughness is relative to the wavelength of radiation emitted or reflected. Specular, or shiny, and mirrored surfaces are usually not considered Lambertian.

The corresponding radiance perceived by a viewer of a tilted Lambertian surface is constant with angle. However, the decrease in intensity from the tilt is compensated by the increase in area the viewer perceives so the surface appears to remain at a constant brightness; see equation 7-4. Therefore the Lambertian surface is a perfect diffuser that gives the viewer a constant radiance independent of viewing angle. Additionally, Lambertian surfaces give a radiance that is directly proportional to its exitance or:

$$L = \frac{M}{\pi} \qquad (7\text{-}5)$$

where:

L = radiance (w/(cm^2 · sr))
M = radiant exitance (w/cm^2)

This characteristic proves to be convenient in the manipulation of thermal sources because their radiation characteristics are often defined by exitance.[xvii]

For a perfectly diffuse source radiating into a hemisphere, the following relationships hold:

$$M = \pi L = \frac{\pi I}{A} \tag{7-6}$$

$$I = \frac{MA}{\pi} = LA \tag{7-7}$$

$$L = \frac{M}{\pi} = \frac{I}{A} \tag{7-8}$$

where:

M = radiant exitance (w/cm^2)
L = radiance (w/(cm^2 · sr))
I = radiant intensity (w/sr)
A = source area (cm^2)

While it appears that the radiance, L, of a Lambertian flat plane should be divided by 2π since it radiates into half a hemisphere, recall that a Lambertian surface does not emit uniformly in direction as the intensity falls off by the cosine. The hemisphere collects less power for angles away from the surface normal. Lambertian surface properties are shown in figure 7-12a and 7-12b.

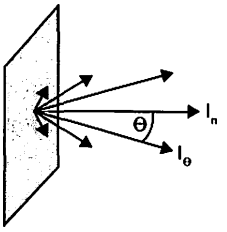


FIGURE 7-12a *A Lambertian surface radiates into half a hemisphere; intensity falls off by cosine.*

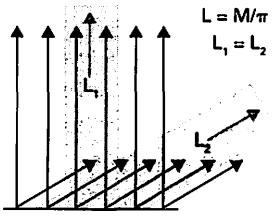


FIGURE 7-12b *Radiance from a Lambertian surface*

Flux Collection

The total power radiated from a source is usually not a concern for a weapon system, because an optical system or a detector can collect only the power within a limited solid angle. Therefore, flux collection is usually concerned with the limited solid angle of the object as seen by the detecting sensor; see figure 7-13.

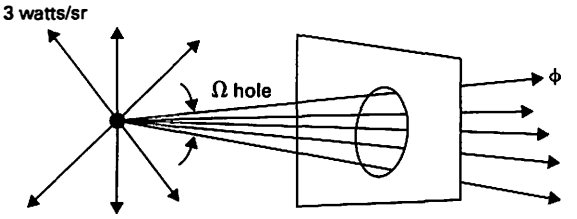


FIGURE 7-13 *Power collected through an aperture*

EXAMPLE 7-4: How much power is radiated through a circular hole in a screen of radius 1 millimeter that is 3 meters from a point source of intensity 3 w/sr?

SOLUTION: The hole, as seen by the point source:

$$\Omega = \frac{A}{r^2} = \frac{\pi(1\times10^{-3}\,\text{m})^2}{(3\,\text{m})^2}$$
$$= 3.5\times10^{-4}\,\text{sr}$$

Proceed to calculate flux (power):

$$\Phi = I\times\Omega$$
$$= (3\,\text{w/sr})\times(3.5\times10^{-4}\,\text{sr})^{\text{xviii}}$$
$$= 1.05\times10^{-3}\,\text{w}$$

Note the difference between the total power radiated by the source, 12πw, as calculated in example 7-3, versus the power collected by the detecting aperture, 1.05×10^{-3} w.

Inverse Square law of Radiation (Spreading)

A common rule of thumb for calculating irradiance on a sensor is the familiar inverse square law of radiation from a point source. By considering a source to be sufficiently small to approximate a point (like many targets at long ranges are) and considering that source to radiate uniformly in all directions (which most targets do not but can be assumed to do so over some angle), this is an excellent approximation.

The implication of the inverse square law is that irradiance is determined by dividing the target intensity (w/sr) by the range squared. Note that this applies strictly to point sources. The development of this "law" is illustrated in figure 7-14.

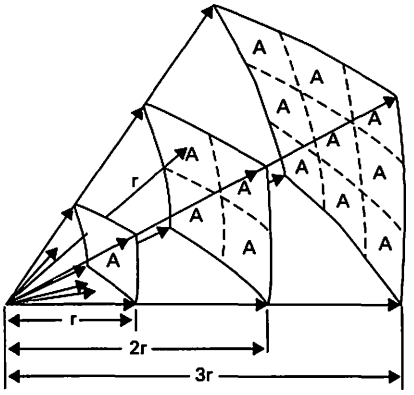


FIGURE 7-14 *Illustration of the inverse square law (spreading)*

It is important to note that the flux leaving a point source within any solid angle is distributed over increasingly larger areas, producing an irradiance that decreases inversely with the square of the distance; see figure 7-15.

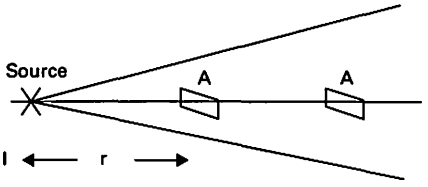


FIGURE 7-15 *Illustration of the inverse square law: Irradiance on the aperture decreases as it moves farther away from the source.*

The following relationship is used to calculate the irradiance of a point source on a spherical surface surrounding the point.

$$E = \frac{\Phi}{A} = \frac{4\pi I}{4\pi r^2} = \frac{I}{r^2} \qquad (7\text{-}9)$$

where:

E = irradiance (w/cm^2)
Φ = radiant flux or power (w)
A = surface area of the surrounding sphere (cm^2)
I = intensity of a point source (w/sr)
r = radius of the sphere surrounding the point source (cm)

Source Geometry

For engineering simplification, electro-optical and infrared sources are classified as either Lambertian sources or point sources, depending on the relationship between their size, their distance from the detector, and the dimensions of the detector aperture. As discussed earlier, a Lambertian source is one that produces an image at the detector optics that is larger than the detector itself or is "large" compared to the detector's field of view. A true point source, though imaginary, is assumed when the range to the source is greater than ten times the largest dimension of the source; a point source is "small" in the detector's field of view.

Electro-Optical Sources—Targets and Background

Sources of Electro-Optical Radiation

Every object in the universe is constantly receiving and emitting EO radiation. The rate of radiation exchange between different objects depends on their temperature and material properties, specifically their surfaces' emissivity and absorptivity. In weapon applications, the concern is primarily those sources that may be considered as targets and the backgrounds against which these targets may be observed. Electro-optical sources of radiation are tactically significant in the following different ways:

1. A source such as a turbojet exhaust, a tank engine, or a person may radiate enough energy to be able to identify the target against a background of differing temperature or material properties.
2. Lasers may serve as an illuminator for an IR weapon's guidance system, a carrier of information in a communication or ranging system, or as a carrier of energy for destructive purposes.
3. Man-made artificial illumination—that is, lights from cities, vehicles, weapons, flares, and so on—may serve as an illuminator of a tactical scene. Targets may contrast enough from this background to enable detection.
4. Natural sources such as the sun, moon, or stars may illuminate a target operating in an EM spectrum that can be amplified to produce imagery or targets with sufficient contrast from the background to enable detection.
5. Conversely, natural and man-made sources such as the sun or other celestial bodies or flares and decoys may serve as background noise against which a detector must compete to identify a target.

Electro-optical signals in general are described by a target-to-background contrast within the bandwidth of the electro-optic sensor. A differential temperature, ΔT, between the target and the background primarily characterizes infrared signals. Even though this quantity sounds like a measured thermal temperature, it is really a measured radiated power quantity. It is this emitted

thermal radiation contrast that infrared sensors are trying to collect.

Visible EO detection, for our purposes, is primarily concentrated on the band of energy that includes visible and near-IR that is reflected off a target. In order to better understand the different sources of radiation, two separate categories will be made: emitters of energy and reflectors of energy. However, remember that EO signatures are a combination of both reflected and emitted energy, and which classification the emitter falls into is a matter of degree.

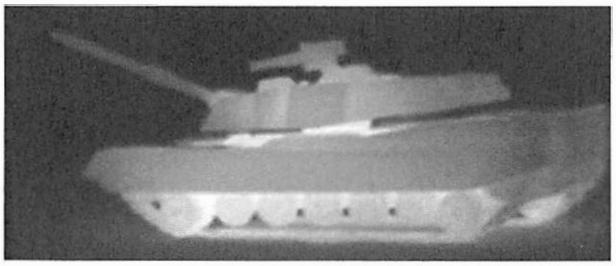


FIGURE 7-16 *Both reflection and emission are combined in most EO signatures.*

Thermal Radiation

Theory

Why do some objects begin to glow when they heat up? Take, for example, an electric stove heating element. Why does it glow in the visible band when it becomes hot? The answer lies within the concept of blackbody radiation. All objects possessing a temperature above absolute zero create EM waves in the infrared region. As the object is heated, its molecules begin to vibrate while held loosely in place by their binding attraction. The vibrating motion is oscillatory and generates EM waves at extremely short wavelengths because of the short intermolecular distances, thus sometimes producing visible radiation.[xix]

As part of the EM spectrum, light is a transverse wave that consists of fluctuations in the electric and magnetic fields of the propagating wave. Like radar, the electric field defines the direction of polarization and activates optical sensors such as photographic film and the human eye.

The foundation of optical radiation theory lies in the idea that radiation is released from bodies in the form of "quanta" or little discrete packets of light called photons. Max Planck first introduced this fundamental law of quantum theory in 1900, and this is a very different way of describing EM waves than that discussed in the previous chapters. All EM waves are quantized—that is, the energy is carried by photons, each having allowable energies of $h\nu$, where h is Planck's constant (6.63×10^{-34} J · s) and ν is the frequency.

$$E = h\nu \tag{7-10}$$

where:

E = energy (J)
h = Planck's constant (6.63×10^{-34} J·s)
ν = frequency (1/s)

Light at a particular wavelength has a specific energy. Figure 7-17 represents the classical chemical explanation of Planck's law in that as electrons move from a higher energy level around an atom to a lower level a photon of light is emitted at a particular wavelength. Planck's equation shows this relationship between the wavelike and particlelike nature of light. This law also shows that longer wavelengths have a lower energy than short wavelengths. Since the EM waves in the RF region are of low relative energy and frequency, the particle nature of the field is of less importance.

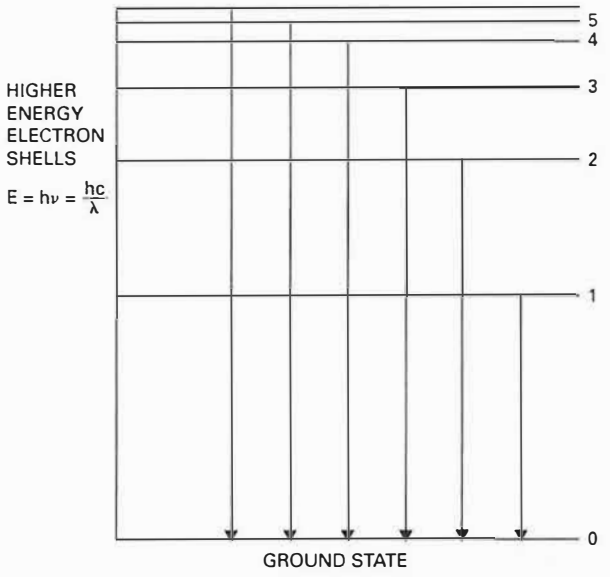


FIGURE 7-17 *Electrons transitioning between states emit photons of energy.*

Therefore, for the RF region, it is the wave characteristic that predominates and the photons need not be considered. For optical radiation, the quantum characteristics of the wave are critical to understanding the theory of EO radiation and detection.

The following examples show how great the amount of energy differs between an EM wave in the RF frequency range and an EM wave in the IR frequency range.

EXAMPLE 7-5:
Energy in a UHF wave

$$E_{RF} = h\nu$$
$$= 6.63\times10^{-34}(\text{w}\cdot\text{s}^2)\cdot 1\times10^{6}(\text{Hz})$$
$$= 6.63\times10^{-28}(\text{w}\cdot\text{s}) = 4.14\times10^{-9}\text{eV}$$

Energy in an IR wave

$$E_{IR} = h\nu$$
$$= 6.63\times10^{-34}(\text{w}\cdot\text{s}^2)\cdot 1\times10^{13}(\text{Hz})$$
$$= 6.63\times10^{-21}(\text{w}\cdot\text{s}) = 0.041\,\text{eV}$$

Note: eV stands for electron volts

$$1\text{ ev} = 1.602\times10^{-19}\text{ J}$$

While the quantities of energy may seem insignificant, the orders of magnitude are vastly different. On an atomic scale, the quantity of energy in the IR example is comparatively large, making it extremely significant in the understanding of EO detectors, operation.

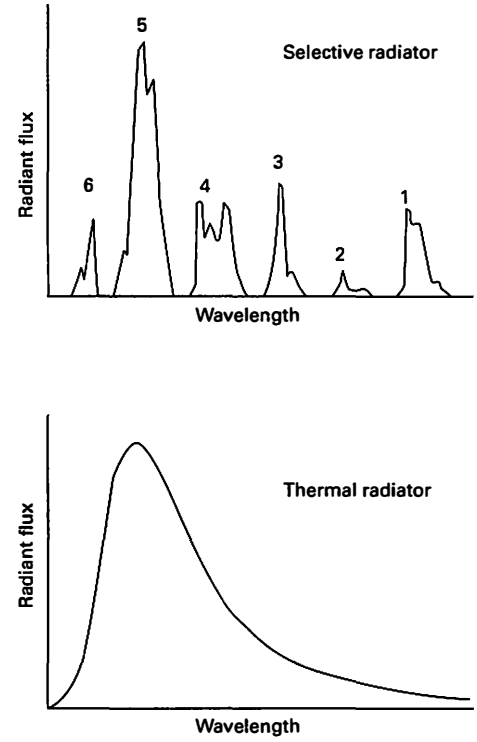


FIGURE 7-18 *Output of thermal and selective radiators*

Thermal vs. Selective Radiators

There are two general types of sources: thermal radiators and selective radiators. Figure 7-18 shows their spectral characteristics.

A thermal source radiates a continuous spectrum of wavelengths with its maximum radiated energy occurring at some particular wavelength. Typical thermal radiators include the hot metal of engine tailpipes (not the gas exhaust), aerodynamically heated surfaces, terrain, personnel, or ships and equipment.

Selective sources have an output concentrated in narrow wavelength bands, called line spectra, caused by the excitation of individual atoms or molecules by various means. A particularly unique selective radiator is a laser that radiates at a nearly single frequency. Its radiation spectrum can be constrained to nearly a single narrow line. Some of the selective radiators that are tactically significant are hot gases from engine exhaust, the shock-excited layer surrounding a reentry body, and reflected laser energy.

Blackbodies

In 1860, Gustav Kirchhoff observed that materials that are good absorbers of energy are also good radiators at any given wavelength. He used the term *blackbody* to describe an object that absorbs all incident radiation and reradiates that absorbed energy with complete efficiency. The term *blackbody* comes from the idea that since it reflects no light, it would appear black to an observer. Thus a blackbody is a theoretical object that is a perfect emitter and absorber of radiation.

Kirchhoff proved that such an object would reradiate its energy in an EM spectrum dependent only on the temperature of the body and not, for example, on its composition. No object is a perfect blackbody because nothing completely reradiates all the energy it absorbs. It is, however, possible to approximate many objects as blackbodies. (A star, for example, acts almost exactly like a blackbody.) Therefore, the blackbody is the standard with which any other source is compared.

Planck's Law of Blackbody Radiation

Planck's law of blackbody radiation describes the exitance at a specific wavelength and temperature (brightness) from a blackbody as:

$$M(\lambda,T) = \frac{2\pi hc^2}{\lambda^5}\frac{1}{[e^{ch/\lambda kT}-1]} \qquad (7\text{-}11)$$

where:

$M(\lambda, T)$ = specific exitance of a blackbody w/m^2-μm
T = temperature in Kelvin (K)
λ = wavelength in micrometers (μm)
h = Planck's constant (6.63×10^{-34} w/sec^2)
c = velocity of light (3×10^8 meters/sec)
k = Boltzmann's constant (1.38×10^{-23} w·sec/K)

Exitance (M) is dependent on wavelength, λ, and temperature, T. Plots of the exitance of blackbodies appear as curves in figure 7-19. Note that total emissive power (the area under the curve) quickly decreases with decreasing temperature.

Experimental study of the spectral distribution of the intensity of blackbody radiation yields two important facts, formulated as the Stefan-Boltzmann Law and Wien's Displacement Law.

Stefan-Boltzmann Law—Total Emissive Power per Unit Area

When Planck's law is integrated over a range of wavelengths, the total emissive power of a blackbody can be determined. The result is the Stefan-Boltzmann Law, which describes the fact that the total emissive power of an ideal blackbody, integrated over all wavelengths, is proportional to the fourth power of the absolute temperature of that body. This relationship is expressed as:

$$M = \sigma T^4 \tag{7-12}$$

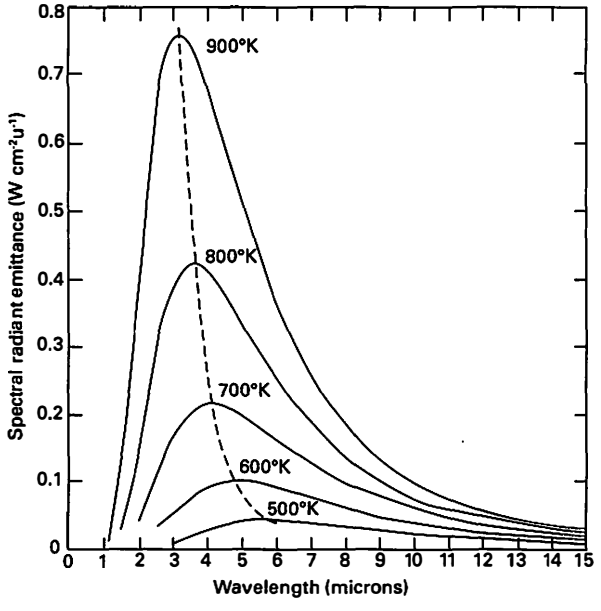


FIGURE 7-19 *Spectral radiant exitance of a blackbody at various temperatures*

where:

M = exitance (w/cm^2)
σ = Stefan-Boltzmann constant (5.67×10^{-12} w/cm^{2}K^4)
T = temperature of the blackbody (K)

The Stefan-Boltzmann Law clearly shows that if a blackbody has a temperature above absolute zero, it will radiate energy. However, to detect a target in a background environment, consideration of the difference or contrast in the radiated powers must be made. Using the Stefan-Boltzmann Law, the difference in exitance from a target and a background can be equated by comparing the difference in temperatures. The law is then expressed as:

$$M = \sigma\left(T^4 - T_e^4\right) \tag{7-13}$$

where:

M = exitance
σ = Stefan-Boltzmann constant (5.67×10^{-12} w/cm^{2}K^4)
T = temperature of the body (K)
T_e = temperature of the environment (K)

Equation 7-13 implies that an object in thermal equilibrium with its surroundings has a net radiated power contrast of zero; therefore a sensor would not be able to distinguish the target from the surrounding environment.

Absorptivity, Reflectivity, and Transmissivity

When searching for a target in a background, it is important to understand that four photonic actions can occur at an object's surface. The quantities that characterize these actions are absorptivity (α), reflectivity (ρ), transmissivity (τ), and emissivity (ε). The definitions of the first three quantities involve the ratio of absorbed, reflected, or transmitted energy to that of incident energy, whereas emissivity (ε) is a ratio of emitted energy of an object to that of an ideal blackbody radiator, which is discussed in detail later. Conservation of energy dictates that:

$$\alpha(\lambda) + \rho(\lambda) + \tau(\lambda) = 1 \tag{7-14}$$

Note that each variable in equation 7-14 is a function of wavelength. The majority of materials are not transmissive in the infrared wavelengths. Some exceptions are zinc selenide and germanium. Most materials that are transmissive in the visible wavelengths such as glass, Plexiglas, and plastic are not transmissive in the infrared.

If a material is opaque, then $\tau(\lambda) = 0$, and:

$$\alpha(\lambda) + \rho(\lambda) = 1 \tag{7-15}$$

Emissivity

The Stefan-Boltzmann Law describes the radiation from a blackbody. However, most sources of radiated energy are imperfect radiators and are called graybodies. To account for the difference in radiation between a blackbody and a graybody, the emissivity (ε) of the object is used. Emissivity is defined as the ratio of the total radiant energy emitted by an object at temperature T to the total energy emitted by an ideal blackbody at the same temperature, under the same conditions.

$$\varepsilon = \frac{M_{\text{source @ temp T}}}{M_{\text{blackbody @ temp T}}} \tag{7-16}$$

The numeric value of an object's emissivity lies between the limits of zero for a completely nonradiating source and unity ($\varepsilon = 1$) for an ideal blackbody. All bodies emit radiation as a result of thermal agitation of their molecules or atoms, whether there are other causes of excitation or not.

The value of emissivity for a given object is a function of the type of material and its surface finish. In general, most nonmetal objects have a very high emissivity, whereas metals have a low emissivity. The relationship between emissivity (ε), absorptivity (α), and reflectivity (ρ), for a given wavelength, is such that for a target at thermal equilibrium, which neither gains nor loses heat, the absorptivity and emissivity are identical:

$$\alpha(\lambda) = \varepsilon(\lambda) \tag{7-17}$$

thus;

$$\varepsilon(\lambda) + \rho(\lambda) = 1 \tag{7-18}$$

From this relationship it is easy to understand that a highly emissive infrared target has a low reflectivity in the same wavelength band.

In the infrared region, the radiant signals come primarily from the target and background emission of blackbody radiation. All bodies above the temperature of absolute zero (zero Kelvin) emit electromagnetic waves. The heat in the body causes molecular vibrations, which in turn cause electron vibrations. These electrical vibrations provide the electromagnetic coupling to produce emission.

A body placed into a vacuum chamber with total darkness, and no impinging radiation, eventually loses all of its heat to the process of blackbody radiation. The amount of radiation that is emitted depends heavily on the target temperature and the surface emissivity characteristics of the target. A hotter target emits more radiation than a cooler target, and a rough target emits more radiation than a smooth target. These characteristics are equally true for background objects as well.[xx]

The radiation from a metal or other opaque material originates within a few microns of the surface; therefore, emissivity of a coated or painted surface is characteristic of the coating rather than the underlying material. However, the visual appearance of material is not always a reliable guide to its emissivity at infrared wavelengths.

For example, fresh snow appears to the eye to be an excellent diffuse, Lambertian reflector, and therefore one might conclude that it has low emissivity. However, the great bulk of the thermal radiation from a body at the temperature of snow occurs above 3μm. The visual estimate of its appearance based on the human eye's sensing radiation centered at 0.5μm is meaningless. Snow actually has an emissivity of about 0.85. Another example is the emissivity of human skin at 32°C, which is close to unity and is independent of its color.

A sample of common material emissivities and reflectivities in the visible spectrum is given in table 7-4.

Since the ideal blackbody is only a theoretical concept, the Stefan-Boltzmann Law must be further refined to account for graybodies as follows:

$$M = \varepsilon\sigma\left(T^4 - T_e^4\right) \tag{7-19}$$

where:

M = exitance
T = temperature of the radiating body (K)
T_e = temperature of the environment (K)
ε = emissivity of the "graybody," for which $0 < \varepsilon < 1$

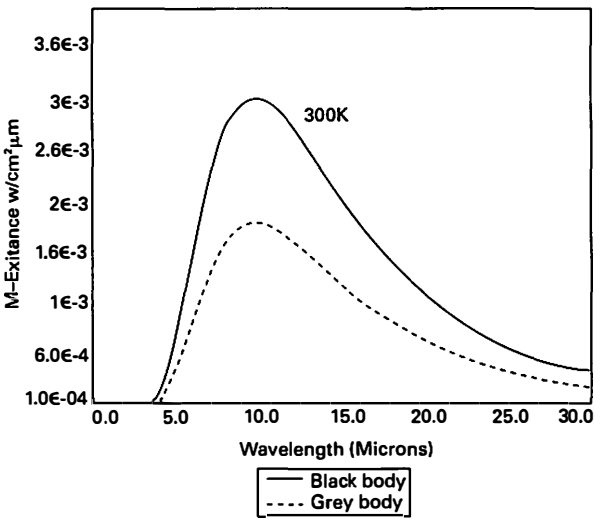


FIGURE 7-20 *Spectral exitance of a blackbody and a graybody at 300°K; note the significant reduction in exitance due to the emissivity.*

Wien's Displacement Law—Wavelength with Peak Radiation

When the derivative of Planck's law, equation 7-11, is taken, the resultant peak wavelength for a given temperature of a body can be determined. This is expressed as Wien's Displacement Law. Wien's Displacement Law states that the wavelength at which the maximum value of intensity occurs varies inversely with the absolute temperature. This is mathematically expressed as:

$$\lambda_{max} = \frac{2{,}898\mu m - K}{T_{Kelvin}} \qquad (7\text{-}20)$$

where:

λ_{max} = wavelength of maximum radiation (μm)
T = temperature of the body (K)

Note that this law locates only the point of maximum radiation and not the radiated energy in other parts of the spectrum.

TABLE 7-4 *Common material emissivity and reflectivity*

Material	Emissivity	Reflectivity
Polished gold	0.02	0.98
Polished silver	0.02	0.98
Polished aluminum	0.06	0.94
Anodized aluminum	0.55	0.45
Polished brass	0.03	0.97
Oxidized brass	0.61	0.39
Polished copper	0.15	0.85
Polished steel	0.07	0.93
Oxidized steel	0.70	0.30
Tin plate	0.07	0.93
Bronze paint	0.80	0.20
Sand	0.90	0.10
Gypsum	0.90	0.10
Paper	0.93	0.07
Brick	0.93	0.07
Rough red rock	0.93	0.07
White lacquer	0.95	0.05
Green or gray paint	0.95	0.05
Lamp black	0.95	0.05
Water	0.96	0.04
Human skin (regardless of color)	0.98	0.02
Graphite	0.98	0.02

Universal Blackbody Curve

When the Stefan-Boltzmann Law and Wien's Displacement Law are applied to blackbody (ε = 1) radiation at a low temperature range of 500°K to 900°K, figure 7-21 is the result. Graybody (ε = 0.66) emission from an object at 900°K is depicted by the broken line. This temperature range includes the general temperatures for the hot metal tailpipes of turbojet aircraft.

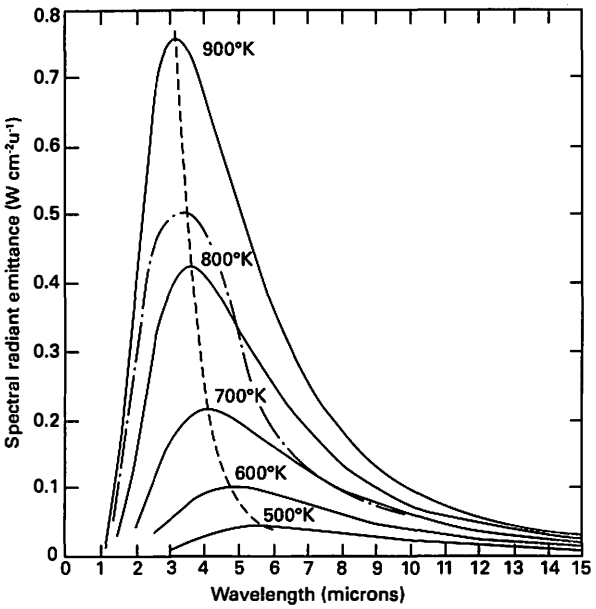


FIGURE 7-21 *Blackbody radiation at lower temperatures (solid lines). Graybody @ 900°K with ε = 0.66 (broken line)*

Several characteristics of thermal radiation are evident from these curves. First, the total output power, which is proportional to the area under the curves, increases rapidly with temperature. Next, the wavelength of maximum radiation shifts toward shorter wavelengths as the temperature increases. Finally, individual curves never cross one another; hence the higher the temperature, the higher the output power density at all wavelengths.

If the same spectral distribution of power is plotted against the product of wavelength λ and Kelvin temperature T, the result is figure 7-22, a universal curve, A, that will suffice at any temperature. In addition, the integral of curve A is plotted as curve B, which provides the fraction of energy radiated below a given value of λT.

For example, at a value of $\lambda T = 4{,}000\mu K$, 48% of the power radiated occurs below this point. This would be true for λ = 8μ and T = 500°K ($\lambda T = 4{,}000\mu K$), λ = 4μ, and T = 1,000°K, and so on. Note that the peak of the A curve occurs where $\lambda T = 2{,}898\ \mu K$, which is the numerical constant in the equation described by Wien's Displacement Law.

Curve B in figure 7-22 is useful in that it graphically determines the fraction of total power radiated by an object that can be detected by an EO sensor of a given

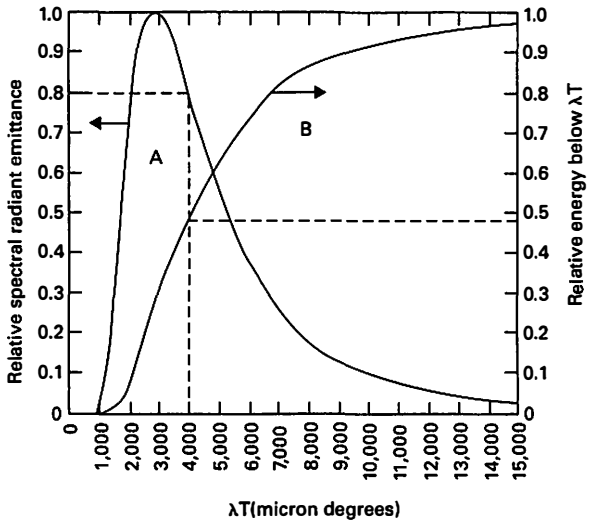


FIGURE 7-22 *Universal blackbody curve*[xxi]

bandwidth sensitivity. From the previous discussion of thermal radiators, it should be apparent that the radiated power per unit area, M, of the Stefan-Boltzmann equation is that power radiated over all wavelengths. Since a detector will respond only to a relatively narrow band of wavelengths, the fraction of the radiated power that is within the bandwidth of a given receiver must be determined. This fraction is termed the bandwidth factor, F, and can be used to make rough approximations of ranges of detection for a sensor, as will be shown in chapter 8.

Selective Radiators

Hot solid objects produce continuous emission spectra or thermal radiation. The oscillations of each atom or molecule are strongly influenced by the adjacent neighbors of each, due to the high density of the substance. However, gases at high temperature and low pressure produce line emission spectra, where each element emits its own characteristic pattern of lines, uninfluenced by adjacent atoms or molecules; this is called selective radiation. As an example, incandescent light bulbs are thermal radiators, whereas neon and mercury vapor bulbs are selective radiators.[xxii]

Figure 7-23 shows how selective radiators have their thermal radiation concentrated in narrow bands of wavelengths.

To have a hydrogen electron excited from its normal energy level (called its ground state) to a higher level, energy must be provided to the atom. The electron will return to its ground state in a very short time, about 10^{-8} sec for most atoms, which is known as the lifetime of the

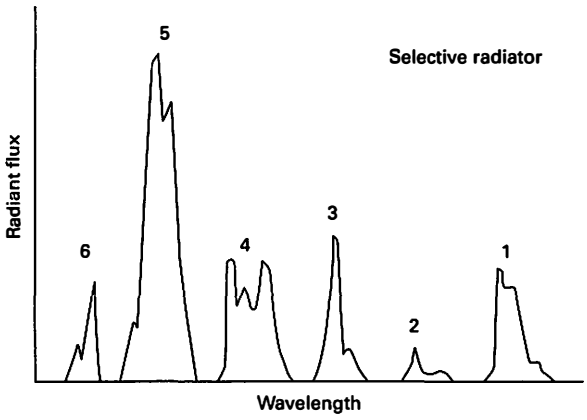


FIGURE 7-23 *Selective radiator*[xxiii]

higher-energy state. In the transition to the lower-energy state, the atom sheds its excess energy in the form of light or other electromagnetic energy.

Quantitatively, Planck's law gives the amount of energy absorbed or emitted by electrons moving from one energy state to another:

$$h\nu = E_1 - E_0 \tag{7-21}$$

where:

E = amount of energy per photon (w)
h = Planck's constant, 6.63×10^{-34} (w/sec^2)
ν = frequency of emitted radiation (Hz)

EXAMPLE 7-6: If a hydrogen atom returns to the ground state from its highest state of excitation, it loses 13.6 ev. Calculate the wavelength of light emitted.

SOLUTION:

$$\lambda = \frac{c}{\nu} = \frac{hc}{h\nu}$$

Substituting this and $E_0 = 0$ for a ground energy state into Planck's law:

$$h\nu = E_1 - E_0$$

Solving for λ:

$$\lambda = \frac{hc}{E_1}$$

Substituting actual values gives

$$\lambda = \frac{6.63 \times 10^{-34}\ \text{w/sec}^2 \times 3 \times 10^8\ \text{m/sec}}{13.6\ \text{ev}}$$

Converting units:

$$1.0 \text{ ev} = 1.6 \times 10^{-19} \text{ Joule} = 1.6 \times 10^{-19} \text{ w/sec}$$

$$\lambda = \frac{1.99 \times 10^{-25} \text{ wm/sec}}{21.8 \times 10^{-19} \text{ w/sec}} = 0.912 \mu\text{m}$$

This ultraviolet wavelength corresponds to the highest energy photon emitted by a hydrogen atom.

Transitions to lower-energy states sometimes take much longer than the normal 10^{-8} second. In some substances these long-lived states may exist for hours or days following exposure to intense light, giving rise to materials that fluoresce or "glow in the dark" while still at low temperatures. CRT screens, aircraft instrument lighting, and the visual detection of submarines by biological luminescence are three applications of fluorescence.

Reflected (External) Radiation

Reflected EO radiation is extremely important for military operations. It is reflected radiation in the visible EM spectrum that allows eyes to detect and differentiate between objects during daylight, and it is the overall decrease in the ambient EO radiation that makes it difficult for humans to tactically operate at night without some electro-optical aid.

Image intensifiers such as night vision goggles are designed specifically to amplify the ambient EM energy within the visible and near-IR spectrum. Since their use is critically important to military operations now and in the future, an introduction to important sources of visible EO radiation and some considerations for image intensification is necessary (see figure 7-24).

Natural Illumination

Solar illumination The sun, of course, is the most significant source of energy. Indeed, for most of humankind's history, the sun has served as an illuminator for our natural sensor system, our eyes. The primary limitation of the sun as an illuminator is that it is not available directly at night, and weather can also drastically attenuate its radiation.

The sun can be approximated as a blackbody at 5,900°K; however, the radiation received on the surface of the earth looks more like selective radiation rather than thermal radiation. This is due to the atmosphere, which is a selective-transmission medium with numerous absorption bands where various wavelengths are attenuated.

The sun's spectrum has a maximum radiated energy at 0.5μm, which, unsurprisingly, is yellow. This wavelength corresponds to the peak of the response of the human eye to light, and the atmosphere provides an excellent transmission medium at these wavelengths. The sun is also a strong source of IR radiation in the near-infrared region (0.7μm to 3μm), and in most of this region the atmosphere is also a good transmission medium; see figure 7-25.[xxiv]

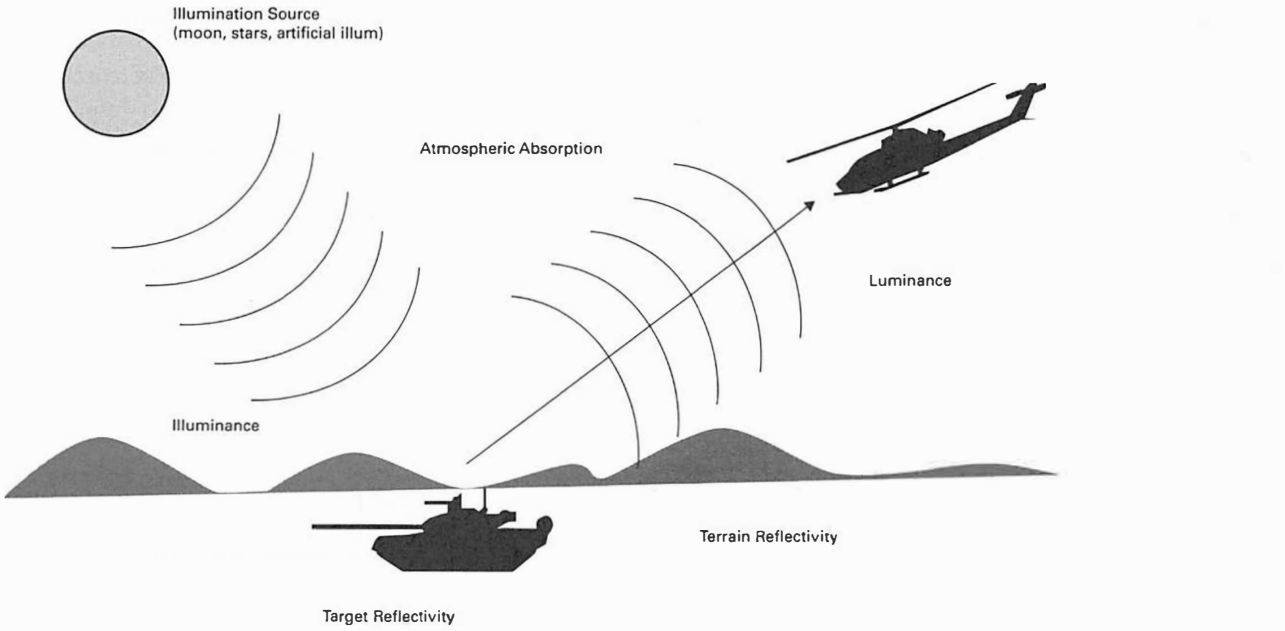


FIGURE 7-24 *Reflected photometric energy is critical for NVG operations*

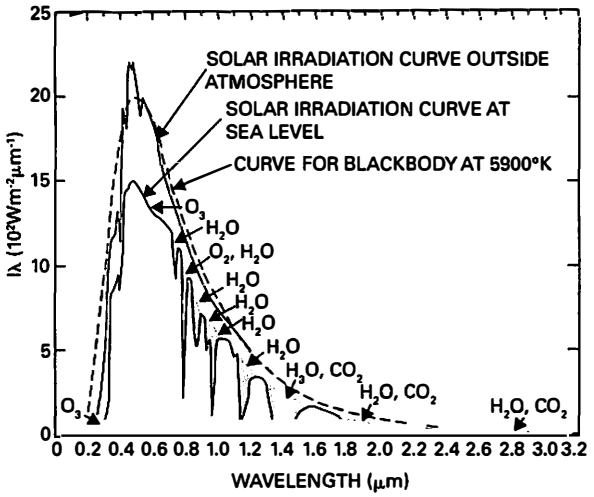


FIGURE 7-25 *Spectral distribution curves related to the sun. The shaded areas indicate absorption at sea level due to the atmospheric constituents shown.*[xxv]

Lunar illumination The moon is usually the primary source of natural illumination for low-level image-intensification operations. The moon reflects about 7% of the sunlight that strikes it. However, the amount of light provided by the moon is highly variable and is influenced by four factors:

1. lunar cycle
2. moon angle
3. lunar albedo (reflectivity)
4. variation in earth-moon distance

The primary lunar illumination factor is lunar cycle or phase—that is new, full, quarter, and so on. Each moon phase provides different levels of illumination, and a lunar month is approximately 29.5 days; see table 7-5.

TABLE 7-5 *Solar and lunar cycle illumination (note that photometric units are used for sun and lunar illumination).*[xxvi]

Sky Condition	Approximate Illumination Lux(lum)
Direct sunlight	$1–1.3 \times 10^5$
Full daylight (not direct sunlight)	$1–2 \times 10^4$
Overcast day	10^3
Very dark day	10^2
Twilight	10
Deep twilight	1
Full moon	10^{-1}
Quarter moon	10^{-2}
Moonless clear night sky	10^{-3}
Moonless overcast night sky	10^{-4}

Moon phases are also influenced by both time of year and global latitude and longitude.

Moon angle—elevation or altitude in relation to the horizon—is the second-most-significant factor that affects lunar illumination. The moon is at its brightest when it is directly overhead and provides less illumination as it rises or sets. Many people will look at a low-angle full moon and assume high illumination; however, this is not the case.

A difference in the albedo (reflectance) of the illuminated portions of the moon surface during the lunar cycle is the third factor. For example, the moon is about 20% brighter during the first quarter (waxing) than it is during the third quarter (waning) due to differences in the lunar surface.

The final and least significant factor is the variation in the earth-moon distance due to the elliptical nature of the lunar orbit around the earth. The changes in illumination resulting from this 26% change in distance are deemed insignificant for EO imaging purposes.[xxvii]

Night sky illumination Moonless nights also have significant usable light for NVG operations. This is due to the large near-IR composition of night sky illumination. This night sky near-IR energy matches the peak sensitivity of many of the newest image intensifiers. On a moonless night, about 40% of the light is provided by emissions from atoms and molecules in the upper atmosphere known as air glow.

Starlight is another significant light source and provides about 0.00022 lux (about 1/10 the level of a quarter moon). In addition, other possible night sky illumination sources include: auroras (luminous rays, ribbons, and arc patterns caused by charged particles); gegenscheins (diffuse, faint light caused by sunlight reflecting off of air particles); zodiac lights (faint elliptical disks around the sun caused by reflection off particulate matter); and noctilucent clouds (usually colored, thin clouds with unknown origin).[xxviii]

Artificial Illumination

An artificial illumination source (i.e., lights from cities, vehicles, weapons, flares, and so on) also provides illumination for NVG operations. As with the sun, these artificial illumination sources can be helpful or hazardous depending upon the relative placement in or out of the NVG field of view (FOV).

Reflectivity (Albedo)

Terrain reflectivity (albedo) greatly influences the overall scene luminance available for NVG operations. Surfaces such as snow reflect more light than surfaces such as asphalt or dark rock and therefore will appear lighter in the

NVG image. The ability to see terrain features with NVGs is solely a function of the amount of light reflected by the terrain. Table 7-6 provides a sample of reflectivity values.

TABLE 7-6 *Terrain/object reflectivity (albedos)*[xxix]

Soils	Dry	Wet	
Dark	0.13	0.05	
Light	0.18	0.10	
Clay	0.23	0.15	
Sand	0.40	0.20	
Surfaces	**Dry**	**Wet**	**Unknown**
Rock	0.35	0.20	
Dirt road	0.25	0.18	
Asphalt			0.10
Concrete			0.30
Desert			0.30
Fields	**Growing**	**Dormant**	**Unknown**
Tall grass	0.18	0.13	0.16
Short grass	0.26	0.19	0.22
Rice	0.12		
Wheat	0.18		
Rye	0.20		
Snow/Ice			**Unspecified**
Fresh snow			0.85
Old snow			0.55
White ice			0.75
Snow and ice			0.65

Terrain contrast is a measure of the difference between the reflectivity of two or more surfaces. The greater the difference in contrast, the easier it is to see terrain or objects. The interaction between terrain contrast and illumination level is critically important for imaging with image intensifiers. Table 7-7 shows some sample terrain contrast comparisons. The ability to distinguish an object from its background is largely a function of this contrast. The primary purpose of camouflage in warfare is to reduce this object-to-background contrast and decrease the detector's ability to acquire the target.

TABLE 7-7 *Sample terrain contrast analysis*[xxx]

Sample Terrain Contrast Comparisons	% Contrast
Asphalt/snow	73
Dirt/grass	41
Sand/leaf	39
Asphalt/grass	18
Grass/leaf	11

Electro-Optic Attenuation

It is a common experience, especially for those living near a large city, to notice that one's view of the city skyline can vary dramatically from day to day depending on the weather—for instance, rain, sleet, fog, or pollution. The same factors that affect the ability to view distant

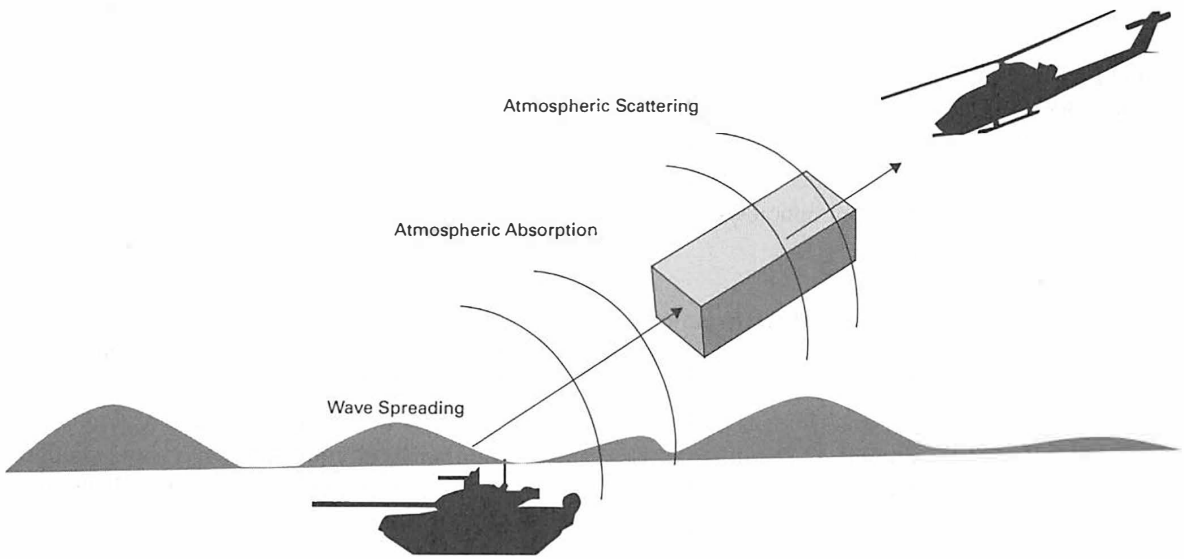


FIGURE 7-26 *Emitted thermal radiation in a typical tactical scenario*

objects also affect electro-optical and infrared sensors. These factors are the result of losses on the radiant intensity.

The radiant intensity and spectral distribution of a source are determined by the thermal radiation properties of its surface for an infrared target or its albedo and contrast from the surrounding environment for a visible target. However, since the radiation leaving the source must travel some distance through a medium, the propagation losses that occur prior to reaching the detector are critically important. The medium in this case is the atmosphere, which is a significant limiting factor to the tactical employment of optical frequencies. Propagation losses are caused by the primary mechanisms of spreading, absorption, scattering, and turbulence.

Spreading and Divergence

The modeling of the spreading from a point source can be done using spherical spreading. The general concept of spherical spreading—waves moving outward in spherical shells—holds true for optics as well as for radar and sound. For a truly spherical wave, the inverse square law applies; that is, the power density (power per unit area) decreases as the square of the inverse of the range. This was derived in equation 7-9.

$$E = \frac{\Phi}{A} = \frac{I\Omega}{A} = \frac{I \times \frac{A}{r^2}}{A}$$
$$= \frac{I}{r^2} = \frac{\Phi}{4\pi r^2}$$

While laser sources may be constructed to have very small divergence, spreading still occurs. Narrowing the divergence of a beam further is still very useful for concentrating power and is equivalent to increasing the antenna gain of a radar system. Just as in the case of radar antenna theory, a narrower divergent beam can be obtained only by increasing the size of the optical lenses or reflectors employed. Thus, lenses or reflectors are the antennas for optical systems.

Atmospheric Structure

The atmosphere is a gaseous envelope that surrounds the earth's surface. It extends for several hundred kilometers above the earth and eventually transitions into the rarefied medium of the solar system. The gas that constitutes the atmosphere is composed of 112 different elements and numerous compounds. The exact composition varies with geographic location and altitude. This is because both atmospheric pressure and temperature change with the vertical structure of the atmosphere.

On a typical day, the morning air temperature starts out cool, gradually warms until it reaches a peak in the afternoon, and finally cools in the evening. Experience shows that the weather is very unpredictable; there are warm fronts, cold fronts, isolated storms, temperature variations, and many other weather phenomena.

Two separate geographic locations within the same general area can have significantly different atmospheric conditions—that is, temperature, pressure, and composition—at the same time. This difference in atmospheric makeup affects atmospheric transmission differently in each region. Despite all these uncertainties, there are certain assumptions that are made, allowing a reasonable prediction of the transmission of radiation through the atmosphere.

The atmosphere consists of elements and compounds. The most common elements in the atmosphere are nitrogen and oxygen, which constitute more than 98% of the atmosphere by volume. Water vapor, carbon dioxide, nitrous oxide, carbon monoxide, and ozone are the major radiation absorbers, with water vapor having a significant effect on the absorption of infrared radiation. Water vapor is also the most variable because of evaporation of water and from bodies of water and damp grounds, and condensation into clouds or dew.

Carbon dioxide does not vary as dramatically as water vapor but does tend to be concentrated around large cities and areas of heavy vegetation. Carbon dioxide is a strong absorber in the midwave infrared (MWIR), however the midwave band is far superior for sensor applications in high humidity conditions than long wave infrared (LWIR).

Earth's atmosphere is generally divided into four layers, based primarily on temperature variations within each layer. These layers are the troposphere, stratosphere, mesosphere, and thermosphere. The temperature in the troposphere decreases with altitude at a rate of approximately 6.5°K/km. Note that between 80% and 85% of the earth's atmospheric mass resides in the troposphere.

The next atmospheric division above the troposphere is the stratosphere, where the temperature is initially constant and then increases until approximately 50 kilometers from the earth's surface.

In the mesosphere, the temperature behaves in a manner similar to that of the troposphere. Temperature decreases with altitude to about 85 kilometers and then strongly rises in the thermosphere. A graph of the temperature versus altitude function is shown in figure 7-27.

In the same manner as temperature, both atmospheric pressure and density vary with altitude. Table 7-8 is a representation of the relationships between altitude, pressure, temperature, and density based on the U.S. Standard

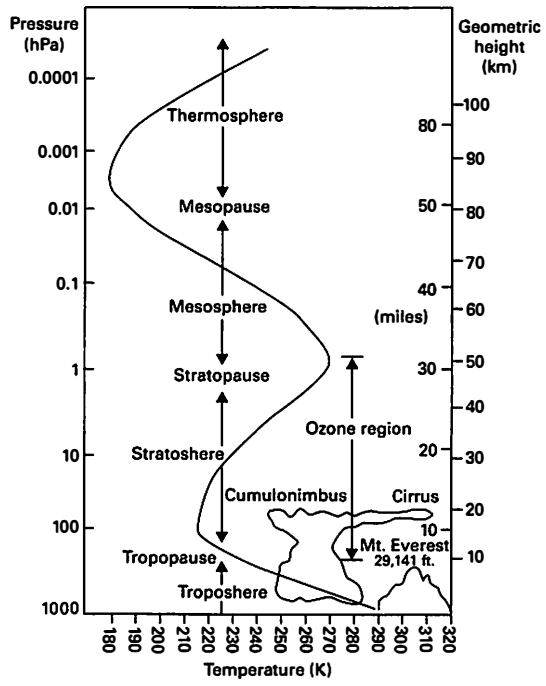


FIGURE 7-27 *Vertical temperature profile of a standard atmosphere*[xxxi]

Atmosphere (1976). The three variables—atmospheric temperature, pressure, and density—must all be taken into account when calculating atmospheric transmission.

The number and behavior of the variables involved in the accurate modeling of the atmospheric effects on the flux leaving the objects of a scene to their arrival at the detector can be extremely complex. However, reasonable estimates can be determined with upper- and lower-error bounds within assumed conditions, as is shown in the following sections.

TABLE 7-8 *Parameters as a function of height. Note: Pressure units hPa stands for hectopascal* (10^2 *PA*).[xxxii]

Height (km)	Pressure (hPa)	Temp. (K)	Density (g/m^3)
0	1.013×10^3	288	1.225×10^3
2	7.950×10^2	275	1.007×10^3
4	6.166×10^2	262	8.194×10^2
6	4.722×10^2	249	6.601×10^2
8	3.565×10^2	236	5.258×10^2
10	2.650×10^2	223	4.135×10^2
12	1.940×10^2	217	3.119×10^2
14	1.417×10^2	217	2.279×10^2
16	1.035×10^2	217	1.665×10^2
18	7.565×10^1	217	1.217×10^2
20	5.529×10^2	217	8.891×10^1

Absorption

Absorption occurs when a photon of radiation is absorbed by a gaseous molecule of the atmosphere that converts the photon into kinetic energy, thereby changing the energy level of the molecule or its component atoms and electrons. This energy change of the gaseous molecule manifests itself as a temperature change.

When examining absorption, it is best to consider the quantum nature of the atmospheric molecule. The energy structure of a molecule is illustrated in figure 7-28. The energy gap E required for absorption is given in Planck's equation, $E = h\nu$ (equation 7-10.) E_g is the energy gap that must be exceeded by a photon for an electron to change energy states within a molecule.

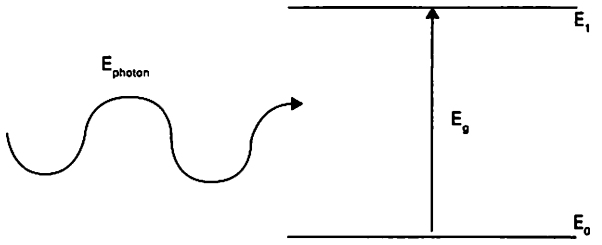


FIGURE 7-28 *Absorption of a photon*[xxxiii]

There is a large number of "allowed" energy state changes within a molecular energy structure, so its absorption properties contain a spectral dependence. For a photon to be absorbed by a molecule, it must be at a sufficient wavelength such that $E_{photon} \geq E_g$ holds true.

This energy difference is related to the different rotational and vibrational states of the particular atmospheric molecules. If the appropriate molecular energy band does not exist, the corresponding photons at that particular wavelength will not interact, to first order approximation with that particular atmospheric molecule.

Because of the quantum mechanical makeup of atmospheric gases, there is no significant absorption or radiation in the visible wavelengths with the exception for H_2O absorption between 0.65μm and 0.85μm. When in the infrared region of the spectrum, however, a number of molecules present are highly absorptive.

EXAMPLE 7-7[xxxiv]: A molecule is sensitive to photons with energies equal to, or slightly greater than, its molecular bandgaps. Lower-energy photons pass and are not absorbed. Determine the wavelength with the most sensitivity (absorption) for a molecule with a gap at 0.18 eV.

SOLUTION: Planck's constant comes in two forms: 6.63×10^{-34} J/sec or 4.14×10^{-15} eV. The speed of light is 3.00×10^{8} m/s. The solution is determined by the relationship $E_g = h\nu$ with E at 0.18 eV, the frequency is 4.35×10^{-13} Hz. The wavelength is then determined from $\lambda = \frac{c}{\nu}$ giving a wavelength of 6.9μm.

Specific atmospheric components that are EO absorptive include ozone, nitrous oxide, carbon monoxide and dioxide, the diatomic molecules nitrogen and oxygen, and water. Transmission plots for primary atmospheric components versus wavelength are shown in figure 7-29; remember that transmission is the opposite of absorption.

Absorption lines or bands are a direct loss process, mostly attributable to the constituents of air, water vapor, carbon dioxide, oxygen, and so on, being cooler than the source of light being absorbed. These constituents have areas of selective absorption, and radiation at these frequencies is absorbed by the molecules and converted into thermal motion, giving rise to an increase in temperature. Figure 7-30 shows the transmission structure of the lower atmosphere.

A weapon designer is thus constrained to work within the available "windows" of relatively high transmission for optimum performance against a selected target signature. An examination of figure 7-30, together with a potential target's emission characteristics, shows several good transmission windows in the target emission bands of interest.

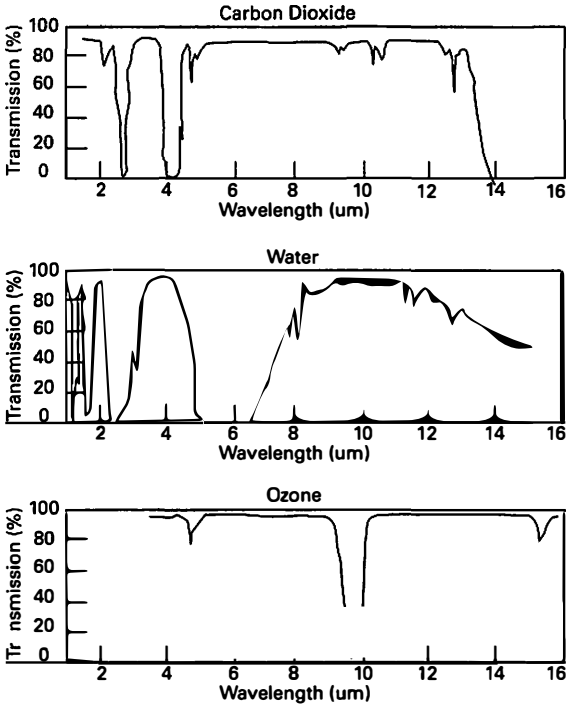


FIGURE 7-29 *Transmission of atmospheric gases for a 1-kilometer path length*[xxxv]

In particular, a good transmission window exists from about 2.9μm to slightly above the 4.4μm region of chemical laser devices. Another good window is located between about 4.4μm and 5.2μm. Finally, from slightly below 8μm to about 14μm is a wide window.

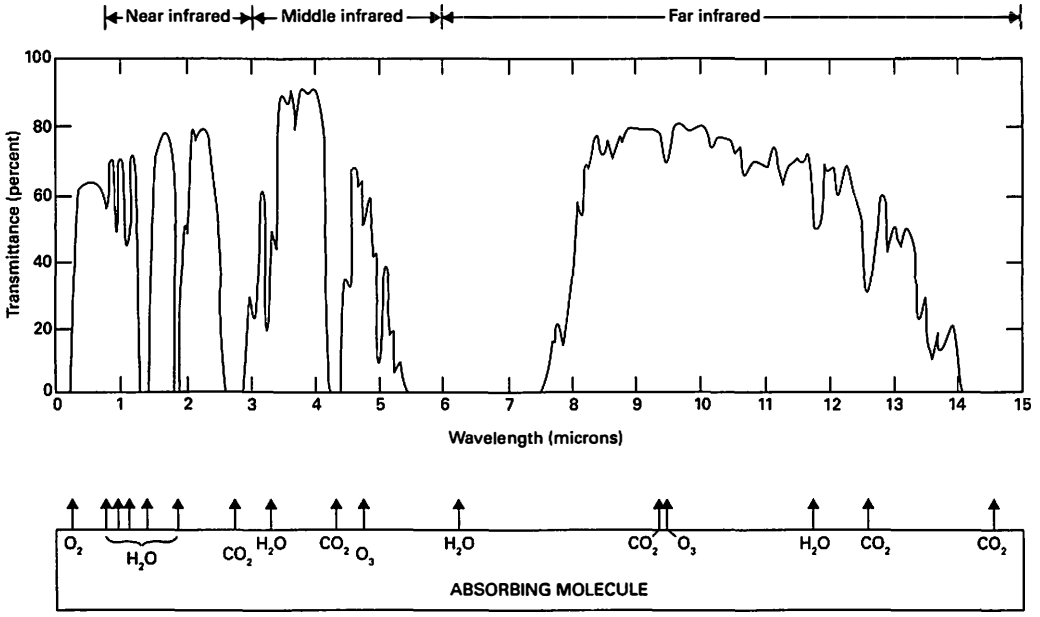


FIGURE 7-30 *Atmospheric transmission windows; transmittance of the atmosphere for a 1,850-meter horizontal path at sea level*

Figures 7-31a and 7-31b show the relative change when one factor, namely humidity, changes in the atmosphere. Notice the severe reduction in the visible band.

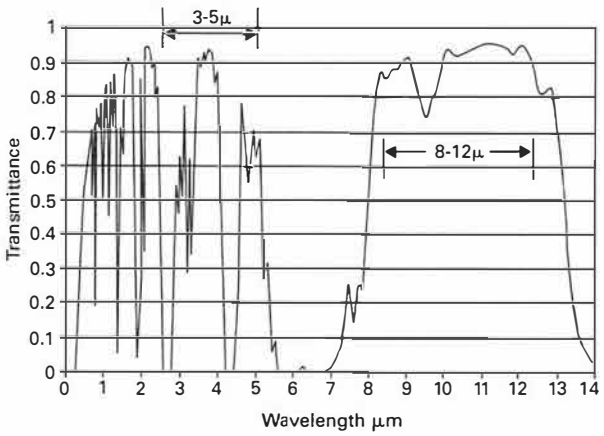


FIGURE 7-31a *Atmospheric transmission for a 45° path from 6 kilometers at 5% humidity*

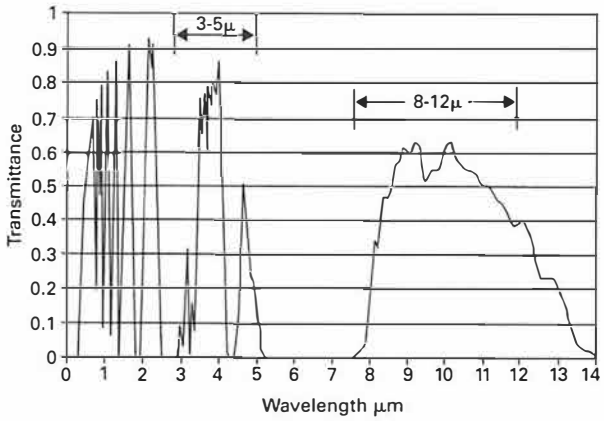


FIGURE 7-31b *Atmospheric transmission for a 45° path from 6 kilometers at 85% humidity*

Scattering

Scattering is the redirection of an object's radiation by particles in the air. Examples of these are water droplets, snowflakes, smoke, airborne dust, pollution, and other human-made and natural aerosols. Scattering causes a redistribution of light rather than a true loss. However, it is certainly a loss for light traveling in a particular direction—for example, from source to detector—and is thus significant as a direct attenuation factor. But the energy lost from the line-of-sight is sent off in other directions rather than being lost as heat.

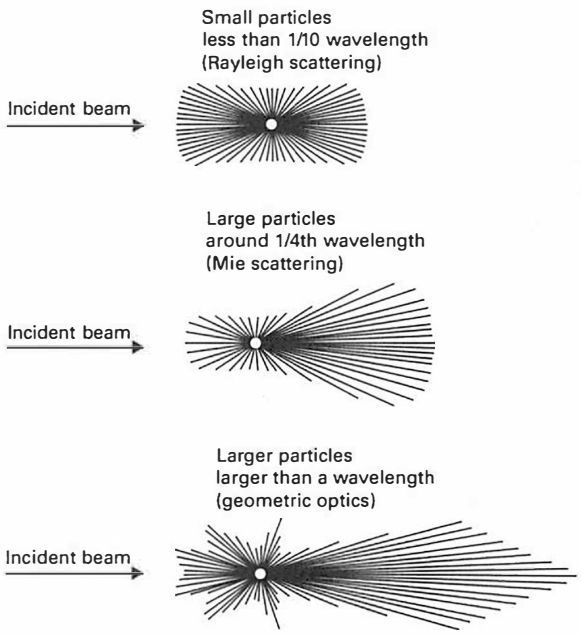


FIGURE 7-32 *Scattering by particles*[xxxvi]

As far as attenuation goes, scattering loss is, in principle, indistinguishable from absorption. One reason for considering it a separate mechanism is that it is not as selective as absorption and causes the detection of a signal outside the main beam. Since significant sidelobes in the RF sense do not show up in the optical regime, the major potential for "seeing" a signal outside the main lobe is due to aerosol scattering.

The process of scattering is the result of photons colliding with atmospheric particles whereby the photon energy is reradiated in all directions. The fraction of the photon's energy extracted and the angular pattern in which it is reradiated is a function of the relative atmospheric particle size to that of the incident photon wavelength.

Scattering differs from absorption because the radiant energy scattered remains in the form of radiation, but is redistributed from the change in the direction of propagation incident radiation's. Typical angular scattering patterns are shown in figure 7-32.

The scattering patterns are divided into three categories that are based on the ratio between the particle size and the wavelength of the incident photon. These patterns are Rayleigh scattering, Mie scattering, and geometricoptics. Some atmospheric particles are shown in table 7-9.

TABLE 7-9 *Atmospheric particles*[xxxvii]

Particle Type	Radius (μm)	Density (per cm³)
Air molecules (Rayleigh)	10^{-4}	10^{-4}
Haze particles (Rayleigh)	1×10^{-2}–1.0	10–10^{3}
Fog droplet (Mie)	1–10	10–100
Raindrops (geometric)	10^{2}–10^{4}	10^{-5}–10^{-2}

The first type of scattering is Rayleigh scattering. It is used when the particle radius is approximately less than one-tenth of a wavelength. The scattering coefficient is proportional to λ^{-4}. As shown in figure 7-32, the angular distribution is symmetrical. It is Rayleigh scattering that leads to the blue color of the sky. The blue wavelengths have a strong scattering interaction with the atmosphere. The blue is scattered in all directions, making the sky appear blue, whereas the red end of the visible spectrum provides little scattering. The attenuation of light for Rayleigh can be represented by the first-order approximation, given by the simplification:

$$\Phi(x) = \Phi_0 e^{-\gamma_r x} \tag{7-22}$$

where:

$\Phi(x)$ = radiant flux at distance $x(w)$
γ_r = scattering coefficient (m^{-1})
Φ_0 = source radiant flux (w)

The second type of scattering is Mie scattering or aerosol scattering. This occurs when the wavelength is approximately the same size as the particle. Detailed Mie theory is based on classic electromagnetic equations with continuity conditions of the boundary between the particles and their surroundings.

Most atmospheric models make certain assumptions to simplify the theory. One assumption is that the atmospheric particles can be represented by simple shapes such as cylinders, ellipsoids, and spheres. It is beyond the scope of this book to derive the details of Mie theory.

Instead, a first-order approximation is provided. This type of scattering is concentrated in the forward direction as shown in figure 7-32. The attenuation given by Mie scattering is of the same form as Rayleigh scattering and is expressed by:

$$\Phi(x) = \Phi_0 e^{-\gamma_m x} \tag{7-23}$$

where:

γ_m = aerosol attenuation coefficient (m^{-1})

The third type of scattering occurs when the particles are much larger than the wavelength. An example of this is raindrops. For this case, the light is primarily scattered in the forward direction and can be approximated with geometrical optics.

Wavelengths detected by an 8–12μm FLIR (forward-looking infrared) are about ten times longer than those detected by night vision goggles (NVG), so the impact of the smaller particles such as smoke or haze on FLIR performance is relatively insignificant. The impact of scattering increases as the wavelength decreases as in the 3–5μm systems. FLIR systems can be very effective in conditions that hinder the human eye and NVGs.

Atmospheric particle scattering coefficient sizes are provided in table 7-10. The size of atmospheric particles in relation to the energy wavelength determines their "attenuation" effect.

TABLE 7-10 *Typical scattering coefficients*[xxxviii]

Condition	Scattering Coefficient (km^{-1})
Dense fog	78.2
Moderate fog	7.82–19.6
Thin fog	1.96–3.92
Haze	0.98–1.96
Clear	0.19–0.39
Very clear	0.078–0.19

Turbulence

Turbulence is a term used to describe time-varying temperature inhomogeneities of the atmosphere. The temperature and pressure variations result in an index of refraction inhomogeneities. The index of refraction variations causes the direction of light propagation to "bend" in various directions. These directional changes vary with time. Turbulence is responsible for a variety of effects such as temporal intensity fluctuations called scintillations. An example of scintillations is the twinkling of stars because star light is a small source transmitting a long atmospheric path length.

Turbulence affects not only visible sensors but also infrared sensors. An example of infrared turbulence can be seen when viewing a distant object across a parking lot in the middle of the summer. The sun heats the pavement, causing the air above it to become heated. This heated air is unstable, and radiation passing through it deviates from its original path. This phenomenon causes image motion, which blurs the image. It is especially noticeable when viewing small, distant objects.

A complete development of turbulence is beyond the scope of this text. The effect of turbulence in EO systems

is most pronounced for coherent systems like lasers, where the propagation of the optical wave front is most critical. Strong turbulence also affects incoherent imaging systems because of the blurring and smearing of images due to the loss of high-frequency information. Although good progress has been seen in the modeling of atmospheric turbulence, there are no accepted standard turbulence models to date.[xxix]

Atmospheric Transmission

The calculation of optical transmission from an object to a sensor involves the three primary processes that affect radiation: absorption, scattering, and refractive index fluctuations (turbulence). The atmospheric components, elements, and compounds discussed earlier are related to absorption and scattering, whereas the atmosphere's structural variations of temperature, pressure, and density contribute to turbulence. The effect of these three factors is both a reduction in the signal amplitude that reaches the sensor from the target and an atmospheric blurring of the image.

The first two processes, absorption and scattering, are usually grouped together under the topic of extinction. Extinction is defined as "the reduction or attenuation in the amount of radiation passing through the atmosphere." It differs from turbulence, where blurring is not uniform across the image.

Overall, extinction affects the entire image in a uniform manner. To understand how extinction affects the transmission of radiation through the atmosphere, visualize a single wavelength of incident radiation passing through an infinitesimal distance *dx*. The change in flux is:

$$d\Phi = -\beta dx \tag{7-24}$$

where:

Φ = radiant flux (w)
β = extinction coefficient (km^{-1})
x = distance traveled (m) as shown in figure 7-33

The negative sign of equation 7-24 indicates the reduction of power, or radiation, from the absorption and scattering processes. In general, the extinction coefficient is made up of two components:

$$\beta = a + \gamma \tag{7-25}$$

where:

a = absorption coefficient
γ = scattering coefficient

Both the absorption and scattering coefficients depend on the incident wavelength. Because of the interaction of incident radiation with molecules comprising the medium, the radiation at the output is changed to $\Phi + d\Phi$.

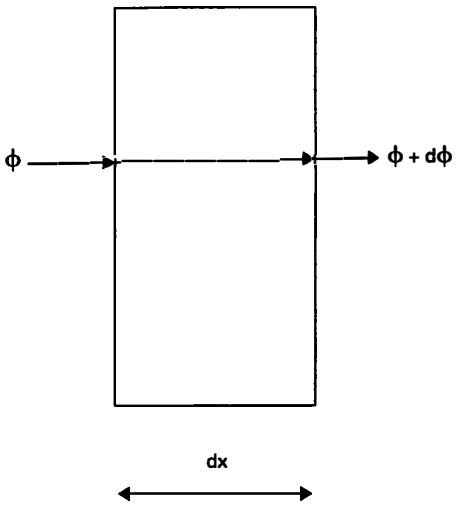


FIGURE 7-33 *Transmission through an infinitesimal distance of atmosphere*[xl]

Equation 7-24 can be integrated to give the attenuation for a finite distance through a homogeneous medium:

$$\Phi = \Phi_0 e^{-\beta R} = \Phi_0 \tau \tag{7-26}$$

where $\tau = e^{-\beta R}$ is the transmittance of the atmosphere over a distance R. This principle is known as the Beer-Lambert law, sometimes shortened to Beer's law. As mentioned above, the extinction coefficient is highly dependent on wavelength, so Beer's law is usually written as:

$$\tau(\lambda) = e^{-\beta(\lambda)R} \tag{7-27}$$

The transmittance of air $\tau_{atm}(\lambda)$ is the product of the molecular extinction transmittance (absorption), τ_t and the aerosol extinction transmittance (scattering): τ_s

$$\tau_{atm}(\lambda) = \tau_a(\lambda)\tau_s(\lambda) \tag{7-28}$$

EXAMPLE 7-9[xli]: A CO_2 laser (10.6-μm wavelength) is propagated through 2-kilometer horizontal path length. Determine the molecular transmittance given that the molecular extinction is 0.385 per kilometer, defined for the given wavelength and relative humidity.

SOLUTION:

$$\tau_m(10.6\mu m) = e^{-(0.383)(2km)} = 0.465 \quad (unitless)$$

If the aerosol extinction is negligible—that is, its transmittance is unity—a 10 w laser provides 4.65 w at a distance of 2 kilometers.

$$\Phi = \Phi_0 \tau_m = 10w(0.465) = 4.65w$$

Ocean Propagation

The ocean is a very highly attenuating medium, as shown in figure 7-34. The only significant "window" is in the visible spectrum, which is frequently called the blue-green window. It is a relatively "dirty" window, however. The attenuation is so great that underwater ranges are severely limited and may be expected to be in tens of meters, with a maximum of 1,000 meters. Thus, optics has very few potential applications underwater. One exception is air-to-subsurface and subsurface-to-air optical communication links.

A second application is active submarine detection. Carried aboard an aircraft, the laser would be pointed downward in a search pattern. The reflected energy from a shallow submarine can therefore be detected.

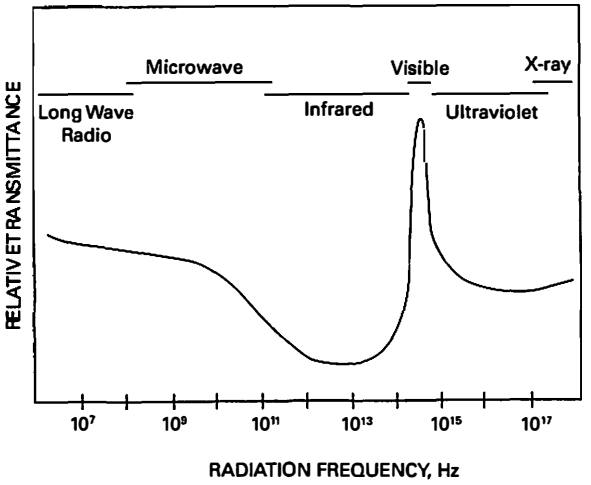


FIGURE 7-34 *Ocean transmission windows*

Propagation in Liquids and Solids

Many liquids and solids are employed in optical systems, both as filters and as lenses. Each nongaseous material has its own unique band absorption spectrum, similar in concept to that depicted above for the ocean. To act as a filter of unwanted noise, a material is selected that exhibits a strong transmission window at the desired wavelengths and minimal transmission elsewhere. Thus, only the desired wavelengths pass through into the detector, and all other unwanted wavelengths are blocked.

When used as lenses, certain crystalline solids such as sapphire, lithium fluoride, and quartz function in the usual optical manner to collect and focus the energy on a detector as well to select a bandwidth of interest.

Although the path lengths that the incoming energy must transit when passing through these various liquids and solids are insignificant when compared to the ranges between the target and detector, the absorption effects are not. This is due to the greater densities of the lens and filter materials compared to the transmitting environment. The intensity of the incoming energy will decrease, roughly, as an exponential function of the depth of penetration and is highly dependent upon the type of material employed.

---

[i] *Encarta*
[ii] Hudson, 464
[iii] Hudson 464
[iv] Hudson, 464
[v] Senich, 217–230
[vi] "Early"
[vii] Driggers, 85
[viii] Hudson, 136
[ix] Palmer
[x] Palmer
[xi] Driggers, 91
[xii] Boreman, 42
[xiii] Hudson, 26
[xiv] Boreman, 113
[xv] Boreman, 44
[xvi] Hudson, 29
[xvii] Driggers, chap. 5
[xviii] Boreman, 45
[xix] Gurvine, 4–6
[xx] Driggers, 102
[xxi] Hudson, 59
[xxii] Gurine, 4–7
[xxiii] Hudson, 32
[xxiv] "Early"
[xxv] Wolfe
[xxvi] "Early"
[xxvii] "Early"
[xxviii] "Early"
[xxix] "Early"
[xxx] "Early"
[xxxi] Butler, 111
[xxxii] Driggers, 130
[xxxiii] Driggers, 133
[xxxiv] Driggers, 133
[xxxv] Driggers, 134
[xxxvi] Driggers, 136
[xxxvii] Driggers, 137
[xxxviii] Driggers, 115
[xxxix] Driggers, 132
[xl] Driggers, 132
[xli] Driggers, 132

8 Electro-Optical Targets and Systems

Introduction

Chapter 7 examined the various characteristics and properties of the EO spectrum. In this chapter, the military applications related to targets and sensors are examined. A target is described as a source because it is emitting radiation that is of interest. As discussed in the previous chapter, there are two basic types of sources, selective and thermal.

A selective source emits most of its energy at one or a few narrow bands of wavelengths. A typical source with this characteristic is the exhaust gases from a jet or rocket engine, but lasers are also in this category.

A thermal source emits broadband radiation over a wide spectrum and has a wavelength of peak power based on temperature, as described in Wien's Displacement Law.

The AIM-9 Sidewinder is an example of a weapon with an EO sensor that guides the missile to the target. The AIM-9 has been continuously upgraded during its 50-plus-year lifetime. Early versions of the Sidewinder had sensors optimized for the wavelengths of typical jet engine exhausts, which required the tailpipe of the target to be directly seen by the sensor (its field of view), limiting attacks to stern-only approaches.

Based on heat signature alone, these early missiles were vulnerable to countermeasures such as flares. As the Sidewinder has advanced with increasingly sophisticated and sensitive sensors, the missile is now considered an "all-aspect" weapon that can detect the target's EO signature from any angle and, through advanced processing, reject countermeasures such as flares.

Targets and Backgrounds

The goal of a sensor is to distinguish the target or source from the background. Sensor response to targets is considered a desired signal, whereas sensor response to the background is considered noise or clutter.

Referring to figure 8-1, it is apparent that a target both emits and reflects radiation. Examples of targets include personnel, buildings, ground vehicles, ships, aircraft, missile casings, and missile plumes. As discussed in chapter 7, a target's radiation is spread, scattered, and absorbed as it propagates through the atmosphere. Therefore, only a small fraction of the radiated energy actually reaches the detector's aperture. Once the radiation reaches the detector, the detector attempts to focus and filter this signal from the noise and clutter to provide useful information.

Noise and clutter are other sources of radiation reaching the detector that are not specifically from the target. Noise includes signals from objects near the target, from the terrain, from the space behind the target, and from atmospheric backgrounds, including clouds or other variances that reduce the detector's ability to discriminate the target signal.

Clutter typically describes the background objects that are similar to the target in size and shape, making the search process and discrimination of targets more difficult.[i] Similar to radar receivers, the sensor itself is also a source of noise because its electronics create internal or "thermal noise" that must be considered as well. Also like radar systems, only when the target signal reaches some minimal level compared to the noise will detection be possible.

Even though the IR region stretches from approximately 1μm to 12μm, only the 3–5μm and 8–12μm regions are used in military applications. This is based on the atmospheric transmission windows shown in the previous chapter and in Wien's Displacement Law. Thermal targets such as personnel and vehicles are normally at temperatures near 27°C, a temperature that corresponds to a peak power around wavelengths of 10μm.

Targets in this temperature range are best viewed by longwave sensors (8–12μm). Rocket plumes, jet exhausts,

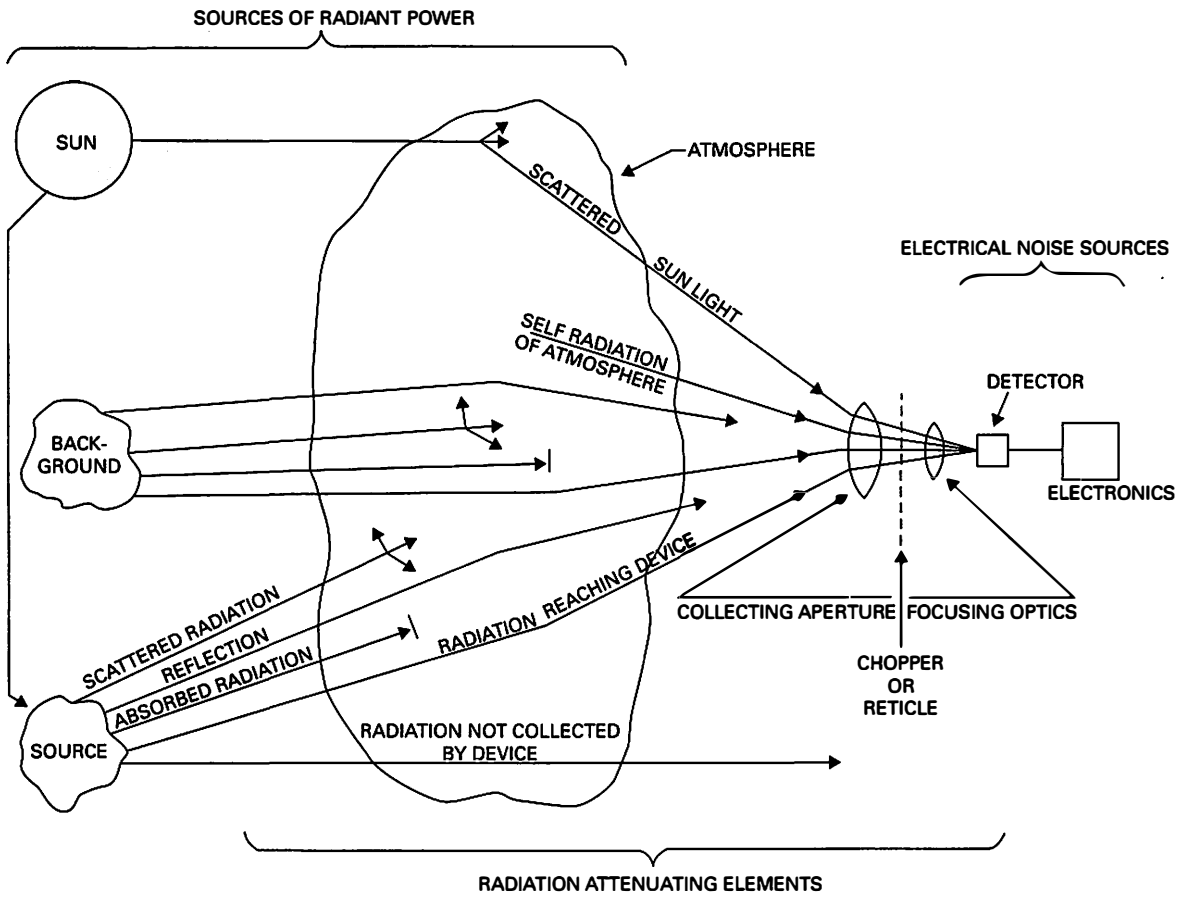


FIGURE 8-1 *Generalized diagram of IR detection*

and projectiles (100–1,000°C) emit peak power near 4μm, and are best viewed by the midwave IR sensors (3–5μm). Remember that all of these targets are still emitting a wide range of wavelengths.[ii]

The thermal image of a commercial airliner shown in figure 8-2 illustrates the importance of both self-emissions and reflections in infrared imagery. In this image, white is hot and black is cold, as is most common. This is termed white-hot polarity. Warm tires show self-emissions as do hot parts near the engine nozzles. However, because of the high reflectivity and low emissivity of the aircraft skin, most of the detected thermal radiation is reflection from the surface.

The lower part of the aircraft appears warm due to reflections from the warm concrete surface below. Reflection of the cold night sky makes the upper part of the aircraft appear colder, with a clear line of demarcation at the horizon. Paint on the aircraft surface provides areas of local low reflectivity. Additionally, the foreground is wet and includes some sky reflection.[iii]

The observed infrared signature from a target, such as the helicopter shown in figure 8-3, also includes both the reflected and the emitted radiation from the object's surface plus the atmospheric attenuation and emission effects. Reflected solar energy is significant in the 3μm to 5μm band and at lower wavelengths. Cold, clear sky reflections are driving the canopy window signatures illustrated in the figure.

Solar illumination is adding to the apparent radiance of the top surfaces. Thermal radiation from the earth, as well as earth-reflected solar radiation, adds to the signature of the sides of the vehicle. Reflections of thermal radiation from hot machine parts and plume emission lines can also be noted on the fuselage skin. Table 8-1 provides a summary of the different radiation sources depicted in both figures 8-2 and 8-3.

Aircraft and Exhaust Plumes

The exhaust gases from an engine contain both water vapor and carbon dioxide molecules in excited states at

TABLE 8-1 *Examples of radiation sources*

Reflection Sources	Emission Sources
Solar	Skin
Sky	Plume (spectral)
Earth	Hot parts
Hot parts off body	
Plume off body	



FIGURE 8-2 *Commercial airliner 8–12μm SPRITE image—white hot polarity*[iv]

high temperatures, thus creating a selective emitter. As seen in figure 8-4, CO_2 has a strong emission line at about 4.3μm. This emission line usually appears broader due to the Doppler effect within the hot exhaust gases.

However, CO_2 in the atmosphere has absorption effects as well. The combination of the broad CO_2 emission line with attenuation by atmospheric gases produces



FIGURE 8-3 *IR signature from a CH-60 Blackhawk*[v]

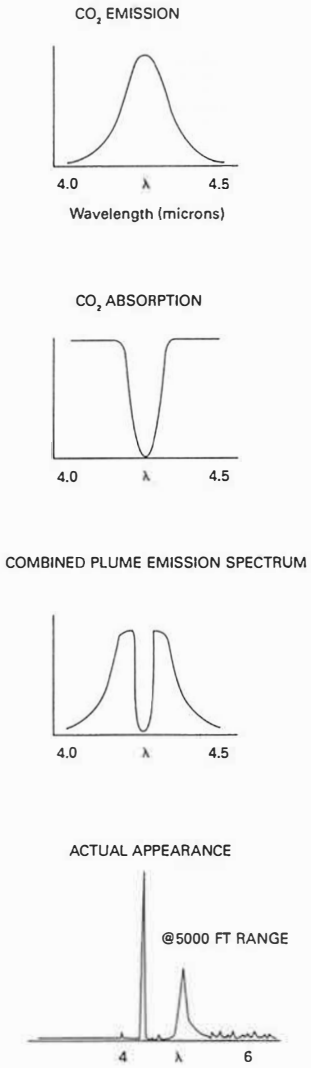


FIGURE 8-4 *Typical exhaust plume spectrum*[vi]

a characteristic dual peak curve as shown. *Note*: Actual emission spectra are much more complex than this simplified diagram suggests.

Jet-powered aircraft signatures are actually a combination of gaseous selective and graybody thermal emissions. Looking at figure 8-5, notice that from a frontal view (0°), the aircraft body obscures the hot engine parts (thermal emissions) and plume core (gaseous emissions); only the outer portions of the plume, together with some small graybody emissions and reflections from the fuselage can be seen.

As target angle increases, greater fuselage graybody emissions and reflections add to the spectrum. At 180° from the nose, the engine exhaust cavity is revealed. This spectrum appears as a thermal radiator except for the "notches" removed by the absorption of atmospheric water vapor and carbon dioxide.

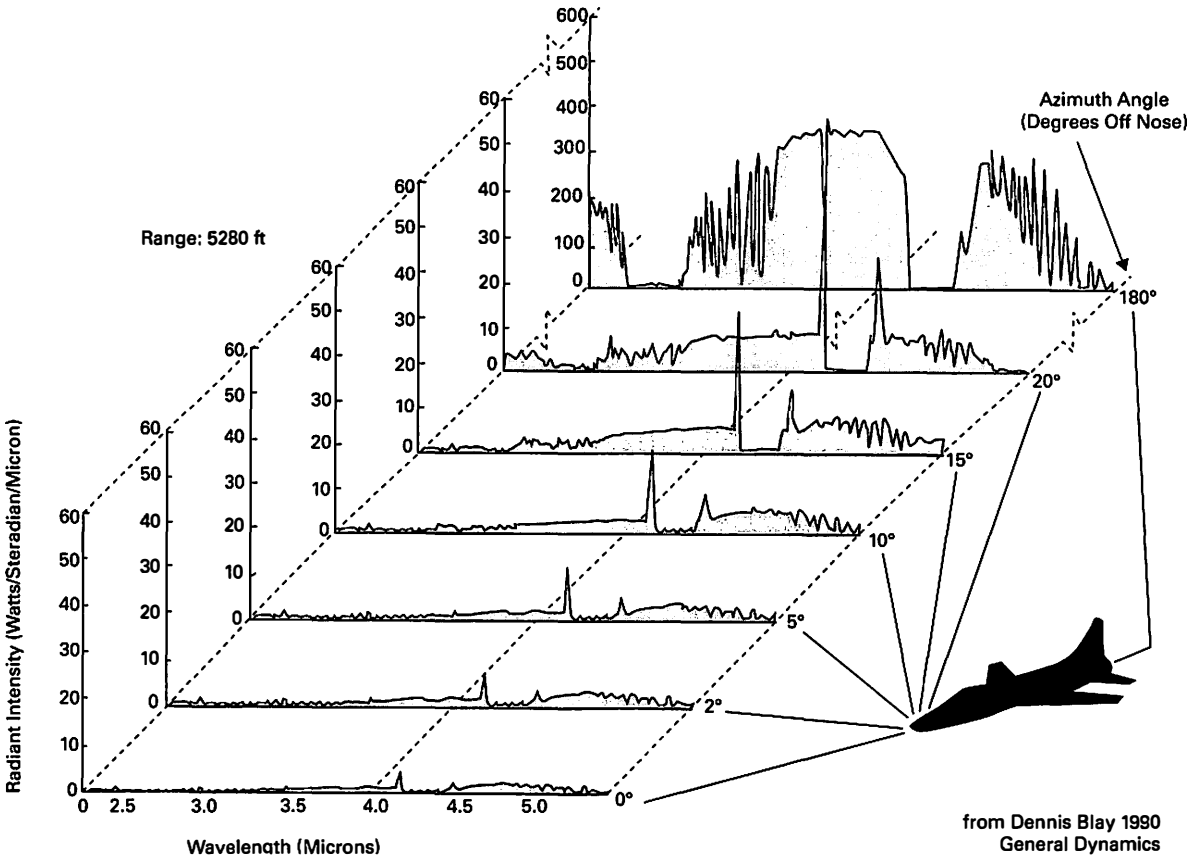


FIGURE 8-5 *Variation in IR signature of a nominal jet aircraft*[vii]

A more in-depth analysis of the principal combustion products, carbon dioxide and water vapor, can be accomplished by examining the spectrum of a Bunsen burner flame. Figure 8-6 provides a qualitative estimate of the spectral distribution of the plume radiation. Note the spikes at 2.8μm and 4.4μm, showing that the radiance of the Bunsen burner is significantly higher at 4.4μm than it is at 2.8μm (by about 3:1). This ratio varies from about 2.5 to 10 depending on the fuel used.

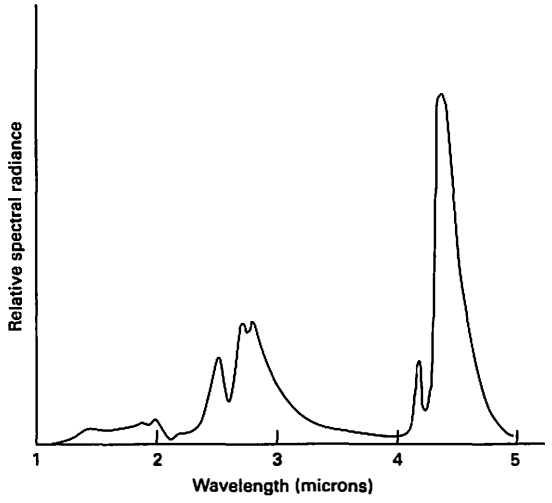


FIGURE 8-6 *Bunsen burner spectral radiance*

This example shows that by analyzing the ratios in the spectral plume, the type of fuel burned can be determined. From the standpoint of detection, the emission band at 4.4μm is likely to be more useful than that at 2.8μm. The band at 4.4μm is also more favorable because of reduced interference from sunlight and reduced atmospheric absorption.

Based on the Bunsen burner model, we view the 4μm emission band of carbon dioxide shown in figure 8-7 in detail. Observing this plot, the solid black curve is the observed radiance in this narrow band. The observed absorption band near 4.25μm is due to carbon dioxide in the atmosphere and in the cool layer of gases surrounding the flame. The dashed curve shows the true shape of the radiant emission of the flame prior to the absorption. The dotted rectangle is an approximation that can be used for estimating the radiance of the plume.

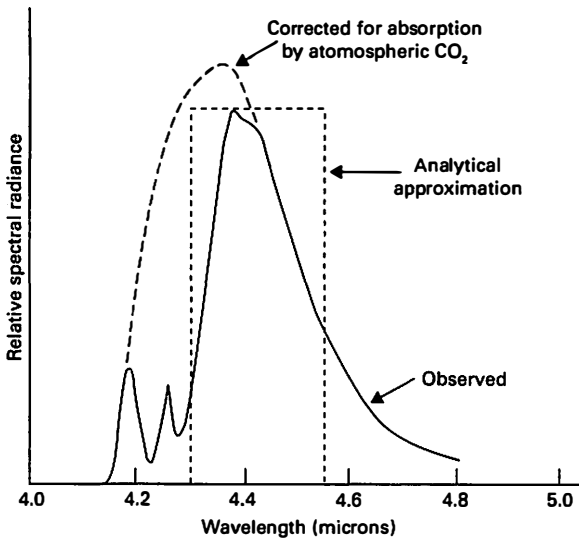


FIGURE 8-7 *4.4 μm emission band*

Figure 8-8 shows the measured temperature for a turbojet and a turbofan engine on a commercial Boeing 707 at maximum thrust at sea level. The size and shape of the exhaust fumes are different for the two types of engines.

A turbofan engine has a large forward compressor that allows a large percentage of air to bypass the later stages of compression and combustion. This excess air is expelled concentrically about the engine so as to form a cold sheath about the exhaust plume. The effect of the fan air in reducing the diameter of the plume is clearly evident.

Another observation is that the lengths of the exhaust plumes are about the same as the length of the aircraft. Therefore the plume, or a portion of it, is visible from almost any aspect angle.[viii]

The signature of a turbojet engine is composed of both the hot tailpipe and the exhaust plume. An estimate of the radiance plume for the two models of the Boeing 707 can be made using the thermodynamic gas laws. The temperature of the exhaust gases leaving the tailpipe of the turbojet is calculated to be approximately 370°C (643°K); similarly, the exhaust gases for the turbofan are about 337°C (610°K). Table 8-2 show the radiance comparison for the plume and the tailpipe of both types of engines.

TABLE 8-2 *Summary of plume calculations for the Boeing 707, altitude 5,000 feet, Mach 0.8, maximum cruise thrust*

For a Single Engine	708-320 (turbojet)	708-320B (turbofan)
Temperature of exhaust gases leaving tailpipe, °C	370	337
Radiance of plume, (4.4μ band only), $Wcm^{-2}sr^{-1}$	0.56×10^{-2}	0.42×10^{-2}
Effective radiance of tailpipe (3.2–4.8μ) $Wcm^{-2}sr^{-1}$	14.4×10^{-2}	10.9×10^{-2}
Ratio of effective tailpipe radiance to plume radiance	25	26

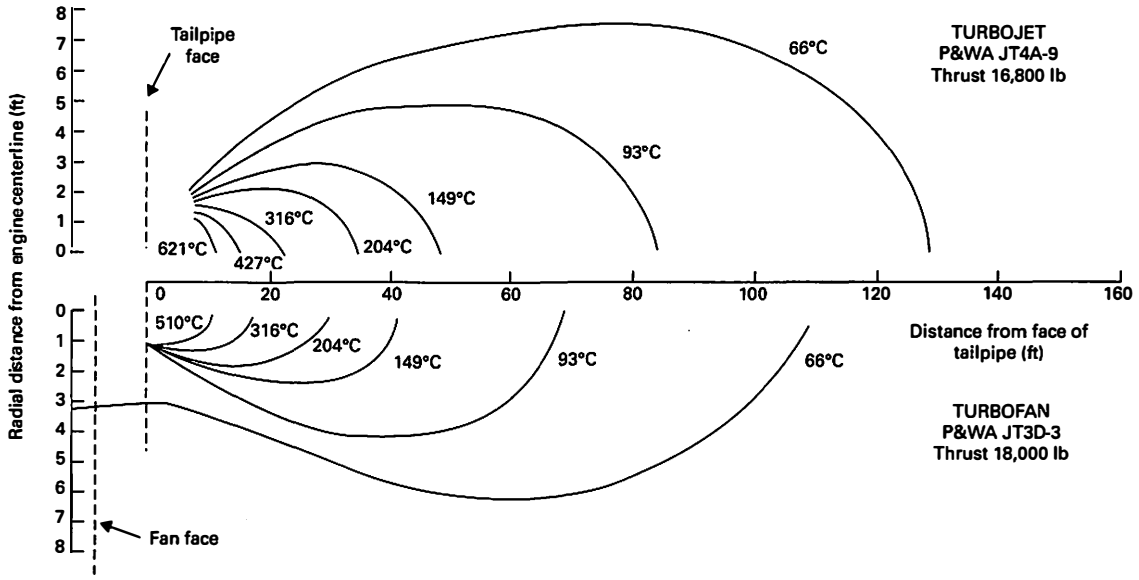


FIGURE 8-8 *Exhaust temperature contours for the turbojet and turbofan engines used on the Boeing 707* [ix]

What is significant is that for both cases, under full power the tailpipe has approximately 25 times more radiance than the plume. This is true even when the value for the plume is calculated for its hottest point—that is, just after the gases leave the exhaust nozzle. The exhaust temperature contours in figure 8-8 show how rapidly plume temperature decreases with increasing distance from the tailpipe.

Farther from the tailpipe, the aircraft engines no longer provide resolved radiance and the whole aircraft acts as a point source. Since the plume has such a smaller amount of radiance than the tailpipe, it is common in the non-afterburning case to ignore the plume in calculating the maximum detection range. However, if the aircraft is viewed from the forward hemisphere or from an aspect angle at which the tailpipes are not visible, the plume is the only source of radiation available for the sensor to acquire.

Modern infrared equipment is sensitive enough to easily detect the Boeing 707 at ranges in excess of 30 kilometers. Accounting for the absorption by the atmosphere, the irradiance for the aircraft is only 7.6×10^{-10} w/cm^2. In comparison, the irradiance at the earth's surface from the sun, for example, is about 0.1 w/cm^{-2}, which is about eight orders of magnitude greater than that calculated for the aircraft.

Aerodynamic Heating Signature

Air vehicles at high speed also have significant IR signatures. Nose cones and leading edges of wings usually have the greatest amount of heating. As shown in figure 8-9, the nose, the leading edge of the wing, and the fairing for the engine have a significant signature. In the right-hand image, the shuttle is about to land and its heat signature is reflecting off the runway.

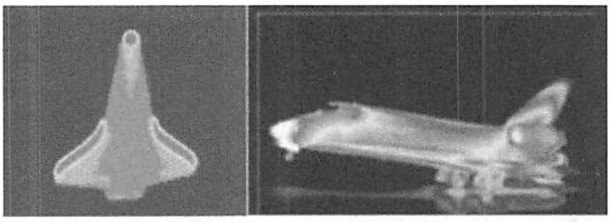


FIGURE 8-9 *Aerodynamic heating of the space shuttle on reentry*[x]

A Cruise Missile Target

Figure 8-10 depicts another possible target, a cruise missile. Its signature can be estimated based on airframe size, propulsion characteristics, and velocity. From its tailpipe, plume, and aerodynamic heating, radiant signatures can be estimated. Other sources such as reflections from the earth can contribute, but they are highly shape and viewing geometry dependent and will not be considered here.

The airframe in figure 8-10 is pictured horizontally at an angle of about 30° off the tail, because this is near optimum for maximizing the total signature.

The cruise missile flies at about Mach 0.7. This corresponds to an aerodynamically heated skin temperature rise

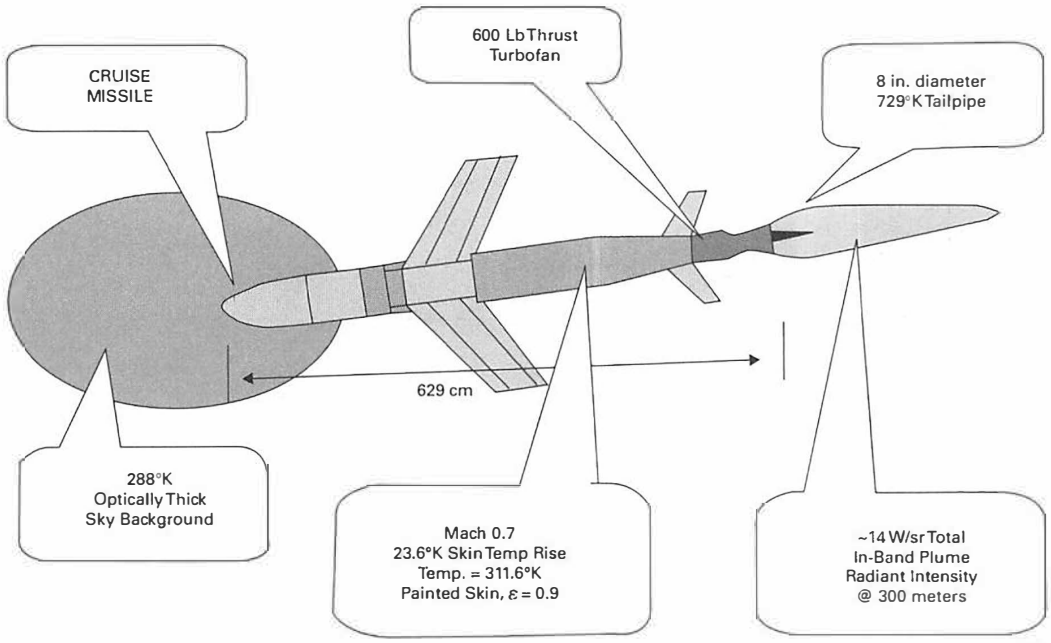


FIGURE 8-10 *A nominal cruise missile*[xi]

of about 24°K in the laminar flow regions. Laminar flow heating is assumed to be uniform over the entire body. Ambient air temperature is taken to be 288°K (15°C), corresponding to standard atmospheric conditions; therefore the missile skin temperature is about 312°K.

Plume and tailpipe signature contributions from the engine are estimated by scaling published data for other engine sizes. The skin and hot tailpipe signatures are computed using the blackbody radiance calculation. Emissivity for the hot tailpipe cavity is assumed to be 1.0, and the skin emissivity is assumed to be 0.9.

Missile Signatures

Figure 8-11 depicts IR imagery in the 3–5μm band of an Atlas rocket. In the IR spectrum, the entire signature is from the missile plume. Any signature due to aerodynamic heating is negligible in comparison to that from the plume.

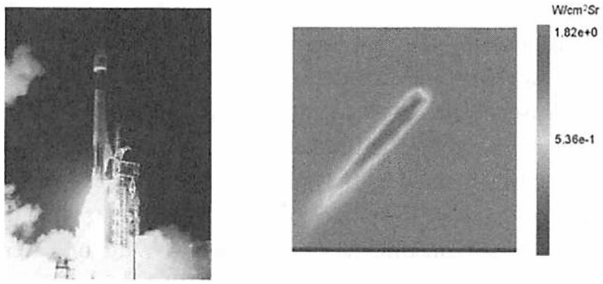


FIGURE 8-11 *MWIR plume image of an Atlas IIA rocket*[xii]

In contrast to the other imagery, this image is a selective emitter, and more detailed analysis of the plume signature could reveal what type of fuel is being used.

An Armored Vehicle Target

The sources of heat in a typical vehicle target can be separated into engine or fuel-burning sources; friction-related sources; and environmental-related sources. Fuel-related sources are exhaust ports, waste heat dissipation areas, and engine compartment covers. They are prominent whenever the vehicle engine is in operation, whether the vehicle is in motion or not.

Some of the fuel-related hot areas of an armored vehicle are clearly visible in figure 8-12. Note the prominent exhaust port on the left side of the vehicle. Other sources of heat include the engine cooling system, air exhaust vent, and engine compartment cover. In the case of moving or recently moving vehicles, wheels, shock absorbers, tracks, and transmission areas all gain heat due to friction.

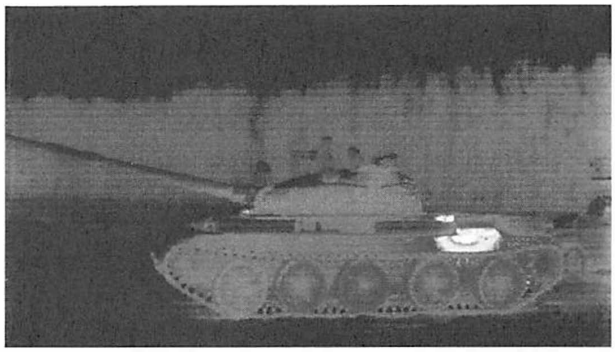


FIGURE 8-12 *T-62 tank signature—side view, 8-12μm band*[xiii]

A rear view of a T-62 tank would show some of the friction-heated areas. Wheel driveshafts would be evident and heat from the engine transmission is also potentially visible from the rear. Other models of tanks with exposed shock absorbers may exhibit high temperatures after traveling over difficult terrain. The upper picture in figure 8-13 shows the rear view of an M-60 tank. The tank treads are warm compared to the tank body due to friction.

Environmental signature sources include sun, wind, rain, and so on. However, it is the effect of solar loading and heat or cold retention by the large thermal mass of the armor that is responsible for the hull and turret signatures seen in figures 8-12 and 8-13.

Figure 8-13 illustrates the role of solar heating and thermal mass in the signature-generation process for heavy armored vehicles. Although the propulsion system and friction are important contributors, solar loading and thermal lag associated with large masses are also major contributors.

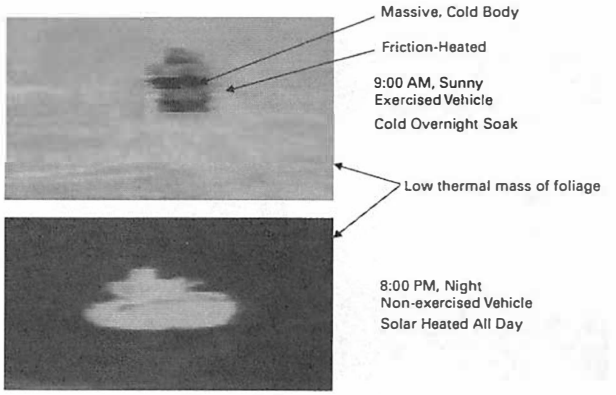


FIGURE 8-13 *Heat sources on armored vehicles—M-60 tank signature, 3–5μm band*[xiv]

The top picture illustrates the effects of cold retention after an all-night cold soak in the mid-40°F temperature range. Though the vehicle is being exercised and both internal heat and tread friction are apparent, cold retention in the armor mass is sufficient to give the hull and turret a large negative contrast. This happens even as solar reflections are acting to reduce that apparent negative contrast. However, the grass background, which has little thermal mass, rapidly heats due to solar loading. The solar-heated background, accompanied by solar scattering, provides a high-radiance background that increases vehicle contrast.

In the lower image, a dormant vehicle has soaked in the sun all day, causing the armor to heat from solar loading. The background grass has rapidly cooled to about air temperature, whereas the large thermal mass of the armor retains heat and produces large positive contrast signature. This signature can be retained for many hours into the night. This phenomenon can be used to considerable advantage as missile seekers can easily spot vehicles in open revetments well after sundown.

Sea-Based Sources

The images in figure 8-14 were recorded with the aid of the Navy AN/SAY-1 camera system, which combines a visible and NIR camera with an LWIR infrared camera. The images illustrate the appearance of ships in various environmental conditions off the coast of San Diego in April 1999.

Turbine engines power these ships and exhibit the same plume and stack emissions as turbine-powered aircraft. Note how the hull contrasts vary significantly depending on day versus night and sunny versus cloudy conditions. Contrast is also affected by the sun angle.

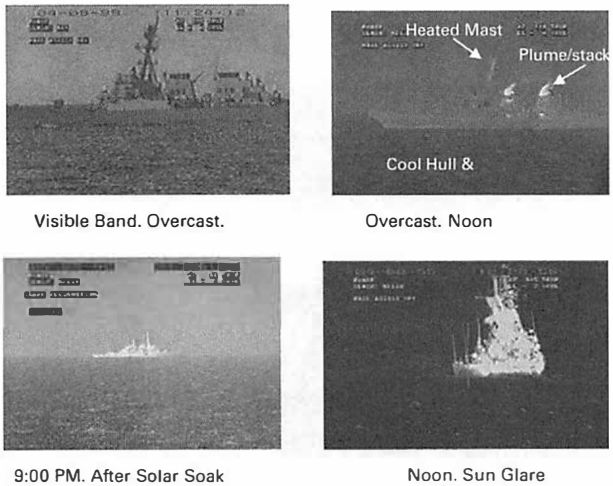


FIGURE 8-14 *Ship signatures, 3–5μm band*[xv]

Solar loading and subsequent heat retention by the large thermal mass is as major a signature driver for ships as it is for ground vehicles.

Lasers

A selective radiator widely used in the military as an optical source is a laser. The word *laser* is an acronym for Light Amplification by Stimulated Emission of Radiation. Lasers emit radiation in that part of the spectrum corresponding to optical radiation, which includes the ultraviolet, visible, and infrared portions of the spectrum.

Laser principles were explored early in the 20th century by the Danish physicist Niels Bohr. His theory showed that atoms emit energy when the electrons orbiting the nucleus move from a higher energy orbit to a lower one. Although electrons may spontaneously change orbits, a laser uses stimulation by an external energy source to direct the orbit change and hence the emission of energy.

Stimulated emission was proposed by Albert Einstein in 1916 and expanded upon by B. A. Fabrikant in 1940. Theodore H. Maiman announced the invention of the first working laser in 1960.

Lasers were first used during the Vietnam War. Their uses have grown significantly since. As a targeting tool, the energy emitted by the laser may be either absorbed by a target, making the target a thermal radiator, or reflected by the target, making the target a selective radiator. Laser's energy may also be modulated, which allows for different lasers of the same frequency to be deconflicted and limits enemy interference.

Weapon systems use integrated lasers such as the Hellfire, Maverick, and Rockeye. Lasers are also used as stand-alone systems as in the Laser Range Finders, Beam Riders, laser target designators (LTD), and the laser spot trackers (LST).

Principles of Operation

Laser is light, a form of electromagnetic radiation. Electromagnetic radiation in the optical spectrum is emitted in the form of a photon whenever a charged particle, an electron, loses energy. Electrons exist in orbits of different levels around the nucleus of an atom. Theoretically, atoms emit photons only when electrons move from a higher-energy orbit to a lower-energy orbit.

For any particular atom, electrons can hold only specific orbits. The lowest orbit held is called a ground state. As the orbit is increased, the amount of energy possessed is also increased. Therefore the greatest energy is released when an electron moves from a very high orbit to a low orbit or its ground state. When the electron is stimulated

TABLE 8-3 *Laser mediums*

Name	Wavelength (nanometers)	Typical Operation
	Solid State Lasers	
Alexandrite	700–830	Pulsed/CW tunable
Neodymium: YAG	1,064.5	Pulsed/CW
Neodymium: glass	1,064	Pulsed
Ruby	694.3	Pulsed
Titanium-sapphire	600–1060	Pulsed/CW tunable
	Gas Lasers	
Helium-neon	543–632.8	CW
Krypton	350–647	CW
Argon	350–514.5	CW
Deuterium fluoride	3,800–4,200	Pulsed/CW
Carbon dioxide	9,000–12,000	Pulsed/CW
	Dye (Liquid) Lasers	
Nitrogen	360–650	Pulsed
Argon	560–640	CW

by energy to a lower state, the emission of photons is controlled, which is highly desired in laser weapon systems.[xvi]

A laser is a simple system of stimulated emission. A lasing medium, a pumping system that supplies energy to the lasing medium, and a resonant optical cavity are necessary. The lasing medium, which may be a solid, liquid, or a gas, determines the specific frequency of the laser. Examples are presented in table 8-3.

The pumping system supplies optical (photon), electrical, chemical, or nuclear energy sources. As the electrons react with the energy source, a few electrons emit photons as they decay spontaneously. A chain reaction starts, emitting photons of precisely the same wavelength, phase, and direction. This occurs in an optical cavity formed by mirrors and lenses that amplify the emission. One of the mirrors is partially reflective, allowing some of the energy to be transmitted through the mirror as a laser beam.

The selection of the type of lasing medium, pumping system, and composition and shape of mirrors is critical in the laser's performance.[xvii]

Like radars, lasers may operate in pulsed or continuous-wave modes. In a CW laser, the pumping system supplies energy continuously, resulting in a continuous laser output. In a pulsed laser, the pumping system is applied in pulses, normally a flashlamp, when using solid lasing mediums (solid-state laser) or electromagnetic energy in a gas medium (gas laser).

Characteristics

Lasers may be described as a function of their coherency, divergence or collimation, efficiency, and lasing medium. Because photons are emitted at specific frequencies, the electromagnetic wave distinguishing a laser emission is characterized by a nearly exact wavelength. This coherent light retains its characteristics over time. At some point, small changes in wavelength will occur resulting from the very small emitted wavelength differences.

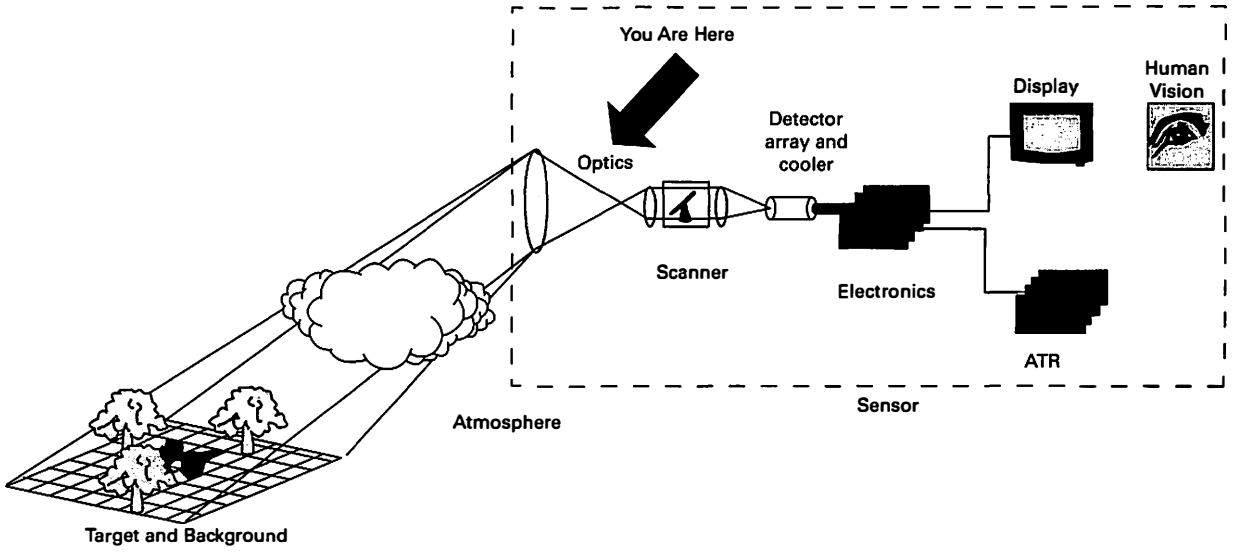


FIGURE 8-15 *Optical characteristics of EO and IR detectors*[xviii]

The time during which the laser wavelength does not change is called the temporal coherency time. The temporal coherency length is the speed of light (3×10^8 m/s) multiplied by the time. This is typically measured in kilometers. The difference in the origin of the electromagnetic wave affects spatial coherency, the change of wavelength over time. Lasers are said to be very coherent, both temporally and spatially. The same emission of EM at nearly exact wavelengths results in a laser being monochromatic—that is, of a single color.

Because there is some small variation of wavelength, it is possible to design lasers that can emit energy at several wavelengths simultaneously. These lasers may be tuned within a band to emit at a single frequency.

Lasers can have very high collimation, which means that the energy dispersion is very low. As the energy travels, high collimation retains the energy in a very narrow beam that slowly diverges from the energy's (laser's) source. The divergence angle is very small, usually a milliradian in size. The area lased on a target can be calculated using this angle.

The process of exciting electrons to a high-energy orbit, and holding them there while adding more energy to stimulate their transition to a low-energy orbit, requires significant amounts of power. Thus, the creation of lasers is inefficient; however, because the laser is coherent, monochromatic, and collimated, the laser is a very powerful weapon over a distance.

Optics

The use of optics is necessary to collect and focus the radiation from the source onto a detector (see figure 8-15). The heart of an optical system is the lens and filter. The lens is simply an optical component consisting of one or more pieces of glass or other substances that are transparent to the wavelength desired. Based on the material's index of refraction EO waves that travel through the lens refract based on Snell's law. Through careful manipulation of the refraction, useful exploitation of the image can occur. The filter is used to obstruct undesirable wavelengths, thus reducing the noise to the system. Note that glass is opaque to IR radiation.

Magnification

Magnifying systems similar to binoculars and telescopes can be thought of as parts of weapon systems. For example, the scope on a rifle or periscopes on a submarine are able to provide highly accurate bearing and elevation range information.

The magnification for a simple, thin lens can be derived by the construction in figure 8-16. The image

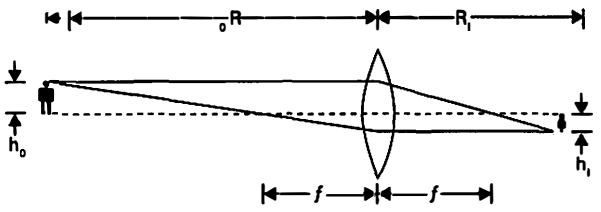


FIGURE 8-16 *Transverse magnification*

equation for a thin lens is:

$$\frac{1}{R_0}+\frac{1}{f}=\frac{1}{R_I} \tag{8-1}$$

where:

R_0 = range to the object from the lens
R_I = range to the image from the lens
f = lens focal length

The sign convention for this equation is as follows: R_0 or R_I to the left of the lens is negative and R_0 or R_I to the right of the lens is positive.

The transverse magnification of the image—which is the factor by which an object's linear size in increased, M_T—is the ratio of the image distance to the object distance and the ratio of the image and object sizes.

$$M_T=\frac{R_I}{R_0}=\frac{h_I}{h_0} \tag{8-2}$$

where:

R_0 = range to the object from the lens
R_I = range to the image from the lens
h_0 = height of the object
h_I = height of the image

Substituting equation 8-1 into 8-2 and simplifying gives:

$$M_T=\frac{f}{R_0+f}$$

For relatively long ranges, where $R_0 >> f$, this simplifies to:

$$M_T=\frac{f}{R_0} \tag{8-3}$$

where:

M_T = transverse magnification
f = focal length of lens
R_0 = range from the object to the lens
R_I = range from the image to the lens

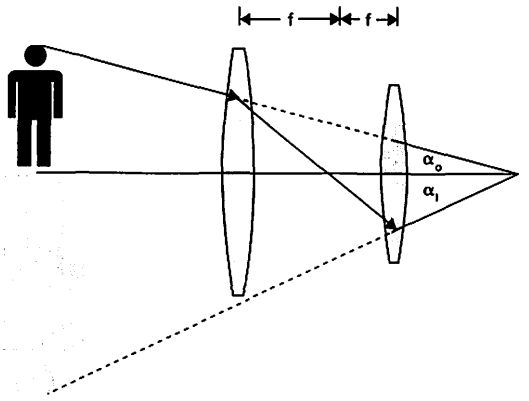


FIGURE 8-17 *Angular magnification*

In most applications, two or more lenses are used to make a telescopic sight, or scope, as shown in figure 8-17. The transverse magnification is not useful in this case. Instead, it is the angular magnification that is relevant. The angular magnificationt—which is the factor by which an object's angular size is increased, M_A—is defined by:

$$M_A = \frac{\alpha_I}{\alpha_0} \tag{8-4}$$

where:

M_A = angular magnification
α_I = angle subtending the aided image
α_0 = angle subtending the object

The correct combination of lenses is achieved by placing the focal point of the objective lens at the focal point of the eyepiece lens. When this is done, the relationship between the focal lengths of the objective and eyepiece lenses will determine the angular magnification. By considering rays coming in from a great distance, the relationship can be constructed as shown in figure 8-18.

Assuming small viewing angles (unlike in the illustration), which is normally the case, the angular magnification is:

$$M_A \approx f_0/f_e \tag{8-5}$$

where:

f_0 = focal length of the objective lens
f_e = focal length of the eyepiece lens

If the field stop of the objective optics is $D_{F.S.}$ and the image size of the object after it passes through the system is D_e then the following relationship can be made:

$$M_A = f_0/f_e = D_{F.S.}/D_e \tag{8-6}$$

where:

$D_{F.S.}$ = objective diameter
D_e = image size of the object

Binoculars and telescopes are specified by the two-number combination: $M_A \times D_{F.S.}$ (in mm). For example, 7 × 50 binoculars have $M_A = 7$ and $D_{F.S.} = 50$ mm. The larger the objective diameter, the better the performance in low-light situations. The field-of-view is controlled by the field stop. Generally, binoculars have a set field-of-view of about 8°.

Field of View (FOV)[xix]

The field of view (FOV) is the range of angles from which the incident radiation is collected by the detector. The field of view may be separated into its horizontal and vertical components, labeled as HFOV and VFOV, respectively. In both cases, the FOV is determined by a combination of the focal length of the lens, f, and the size of the field stop, $D_{F.S.}$. The instantaneous field of view, or IFOV, is the range of incident angles seen by a single detecting element in a focal plane array. The construct definitions are depicted in figure 8-19.

The focal length of a lens is the distance from the center of the lens to the point where all of the incident radiation (or light) coming from a source at infinity is focused. If the source is very far away, the incident rays of radiation will be nearly parallel and are assumed to be so. The lens will refract them all to the same point, namely, the focal point of the lens, as depicted in figure 8-20.

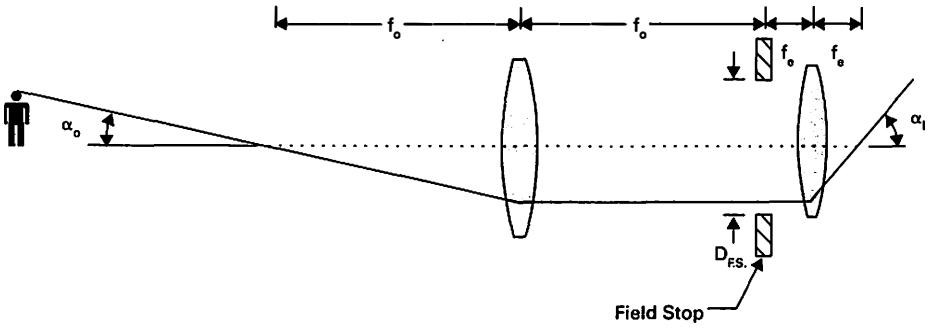


FIGURE 8-18
Magnifying power

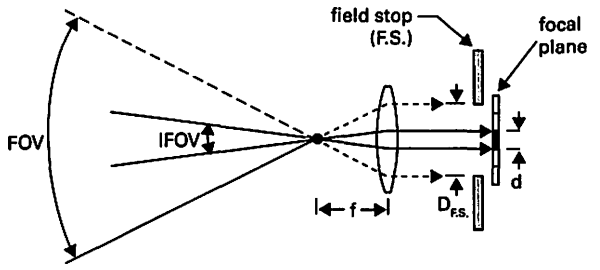


FIGURE 8-19 *Fields of view*

The field stop is a device that blocks rays that are beyond the field stop's dimensions from reaching the detecting element. An example is depicted in figure 8-19. The detecting elements are located at the focal plane, which is usually not in the same location as the focal point. The location of the focal plane determines at what range objects will be brought into focus. The field stop is located just before the focal plane. If there is no physical barrier, then the boundaries of the detecting elements determine the field stop dimensions.

As shown in the geometrical construction in figure 8-19, the diameter of the field stop, $D_{F.S.}$, affects the FOV. If the field stop is made smaller, the FOV will be reduced accordingly. By analogous reasoning, the instantaneous field of view (IFOV) will be affected by the size of the individual detecting element, d. The IFOV and FOV are calculated using trigonometry as follows:

$$\text{IFOV} = 2\tan^{-1}\left(\frac{d}{2f}\right) \tag{8-7}$$

$$\text{FOV} = 2\tan^{-1}\left(\frac{D_{F.S.}}{2f}\right) \tag{8-8}$$

For angles less than 20°, which is generally true for IFOV, the inverse tangent can be accurately approximated by $\tan^{-1}(\theta) \approx \theta$ (radians), in which case IFOV $\approx d/f$ (if $d/f \ll 1$).

EXAMPLE 8-1: Find the FOV and IFOV for a 35 mm digital camera with a 50 mm lens that uses 1,152 × 864 resolution.

SOLUTION: A 50 mm lens has a focal length of 50 millimeters; therefore $f = 50$. Because 35 mm refers to the field stop, $D_{F.S.} = 35$ mm. *Note*: In a film camera, this refers to the size of the film.

$$\text{FOV} = 2\tan^{-1}[35/(2 \times 50)] = 38.6°$$

The instantaneous field of view (horizontal and vertical) is calculated by noting that the FOV must be divided by the number of elements.

$$\text{HIFOV} = \text{FOV}/1{,}152 = 0.03°$$
$$\text{VIFOV} = \text{FOV}/864 = 0.045°$$

Depth of Focus

For optimal performance, the focal plane must be placed so that those objects at the desired range will be in focus. For any object range, all the rays from that object will come together at a unique location beyond the lens. If the object is at a very long range, the rays will come together at the focal point; see figure 8-21.

At shorter ranges, that point gets farther and farther from the lens. The distance of the object in front of the lens, R_0, and the location of the focused image behind the lens, R_I, are related by the image equation, 8-1.

If the focal plane is located so that the object is in perfect focus (meaning the object and image distances satisfy the lens maker's equation), the question becomes: At what range from the object does the image become noticeably unfocused? As the object's distance is changed, its image will become spread out on the focal plane.

When the image becomes so spread out that it overlaps the adjacent detecting elements, the overall image will be distorted. So, the size of the detecting element determines the limits beyond which the image is considered to be unfocused. The range of object distances that satisfy

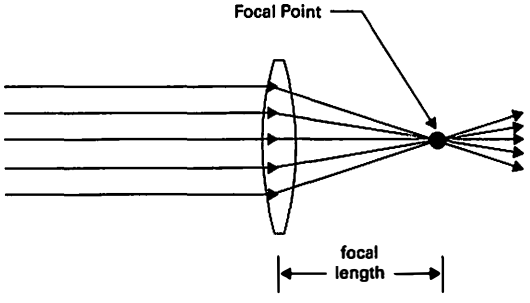


FIGURE 8-20 *Focal point of a lens*

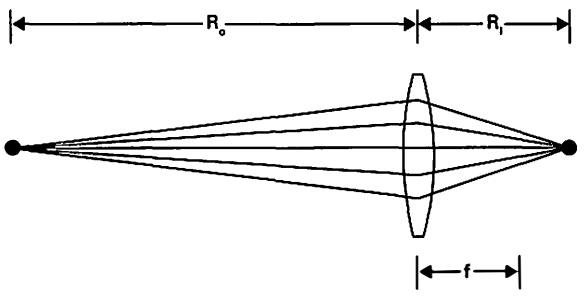


FIGURE 8-21 *Object and image distances*[xx]

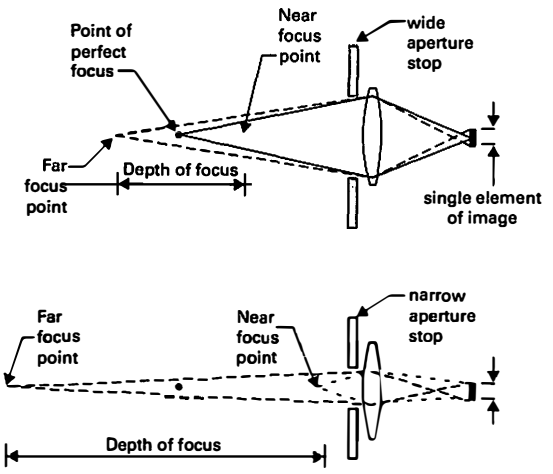


FIGURE 8-22 *Depth of focus*

the criterion for making a suitably focused image is called the depth of focus.

The depth of focus depends initially on the detecting element size and the focal length. In practice, the depth of focus can be controlled. This is accomplished by the use of the field stop. As already explained, the field stop limits the amount of flux that is collected. So it would appear that the largest possible field stop is the most beneficial. However, it must be remembered that the larger the field stop, the shorter the depth of focus, as illustrated in figure 8-22, which is sometimes not as beneficial.

By limiting the range of angles at which the rays may enter the optics, the field stop actually improves the depth of focus. As a trade-off, however, reducing the field stop limits the amount of flux that can be collected by the detecting system, and therefore requires more sensitive detecting elements to achieve the same maximum detection range. This is why smaller apertures require longer exposure times in conventional photography.

Stadimeter Ranging

Range can be obtained from a strictly passive optical system if the size of the object is known. The principle is based on trigonometry. If an object's height, h, is known and the angular extent of the object is θ, then at some unknown range, R, the range can be determined from the following equation.

$$R = h/\tan(\theta) \qquad (8\text{-}9)$$

This result is not particularly useful until the small angle approximation is applied. For small angles, the stadimeter range equation becomes:

$$R = h/\theta \text{ (angle in radians)} \qquad (8\text{-}10)$$

$$R = 60h/\theta \text{ (angle in degrees)} \qquad (8\text{-}11)$$

where:

h = height of object

θ = angle subtended by the object

Equation 8-10 approximates 1 radian as 60°, which enables a good rule of thumb. An object 100 feet tall will subtend 1° at 1 nautical mile. For other heights or angles, the range can be found by taking a ratio.

Another rough approximation of range is determined through the use of field of view and size of the image in the reticle as in the following example.

EXAMPLE 8-2: Using binoculars with an 8° field of view, a 1,200-foot tower is observed that fills one-quarter of the FOV. What is the range to this tower?

SOLUTION: One-quarter of the FOV is 2°. Knowing that a 100-foot tower will subtend 1° at 1 nautical mile, it would subtend 2° at 0.5 nm. The 1,200-foot tower would subtend 2° at 12 × 0.5 nm, or 6 nm; therefore $R = 6$ nm.

Scanners and Detectors

Typical sensor design and analysis revolves around some operational requirement that includes a target under some background conditions. A typical requirement for a military EO system is similar to the following statement:

> Provide a 90% recognition probability of a T-62 tank at a range of 2 kilometers under midlatitude summer weather conditions.

This example gives a system requirement that is based on specific targets and must be translated into parameters that are meaningful to a sensor designer.[xxi]

The detector is the part of the sensor system that transforms the optical signal into an electrical signal (see figure 8-23). What is needed is an electrical output that is proportional to the incident optical power. The detector component in the sensor plays a key role in determining system-level parameters including spectral operating band, sensitivity, and resolution. Additionally, the detector material characteristics and the operating temperature determine the spectral response.

A strong energy band for photon acceptance must also be present for the energy conversion process to occur. Further, detector sensitivity is a function of material, wavelength, detector size, bandwidth, and shielding. Finally, the resolution is set by the detector size and pitch (center-to-center spacing) and the optical focal length. The detector must be distributed over the image plane in order to convert the image into a spatial signal.

The detector distribution across the entire image is accomplished by scanning the image across single detectors

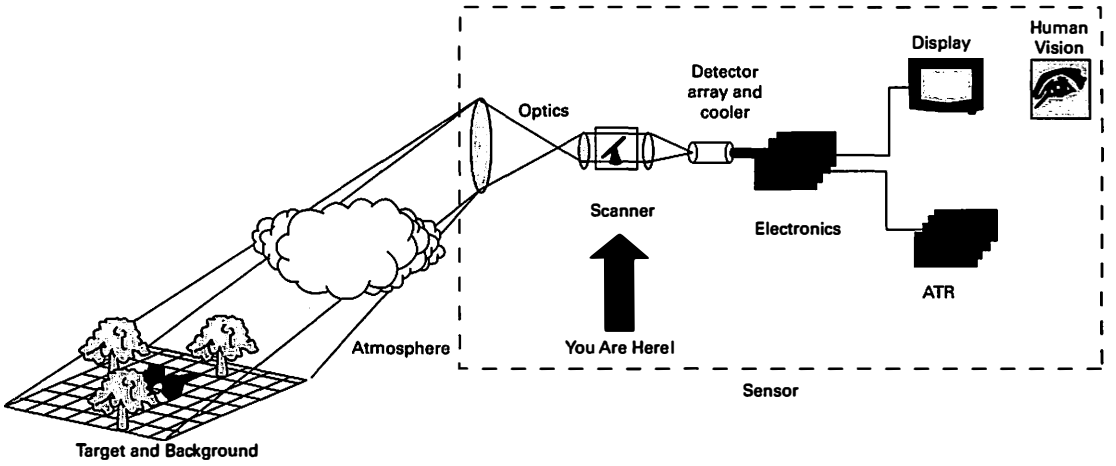


FIGURE 8-23 *The "System"—detectors and scanners*

or linear detector arrays. Older-style detectors like the early Sidewinder missiles used a single-detector element and a spinning mirror or reticle that scanned the field of view. Many missile seekers still use this technique, but the current trend is to use an array of many discrete elements in the focal plane called a "string" array. This is the type of seeker used on the Israeli Python missile and the AIM-9X.

This section of the text focuses on spectral sensitivity and detector resolution along with the necessary scanners required to perform the spatial image representation.

Scanners

The optical detector's FOV is scanned by one of three basic techniques: serial scanning, parallel scanning, or staring array. Each type has its advantages and disadvantages that must be considered when designing an electro-optical system.

Serial Scanning

The serial scanner uses two mirrors to optomechanically scan the detector area across the scene in a series of swaths with each successive scan immediately below the previous swath; see figure 8-24. The horizontal axis of the detector's image is generated as the thermal energy is scanned across the rotating mirror, whereas the vertical is built by the repetitive scans that are adjusted by the oscillating frame mirror.

Because the detector array has multiple (eight, for example) separate detector strips positioned in a vertical line, each swath scans eight lines of the thermal image. The scanning continues until the full FOV is built. This scanning technique is referred to as serial-parallel processing because each detector element is used to scan a scene line by line and element by element, so each element constitutes one line of the thermal image.

There is a space between each detector strip in the array, which could cause a gap in scene information and reduce the resolution of the image in the cockpit display. However, this missing imagery is picked up with a 2:1 interlace feature that is integrated into the scanning process. Interlacing means that after each field is built, the scanner is adjusted to create another field that includes those portions of the field that were previously missed.

These two fields constitute a complete image (frame). This requires rapid scanning, for example, as each field is created at a 60 Hz rate and the entire image (frame) is refreshed at a 30 Hz rate.

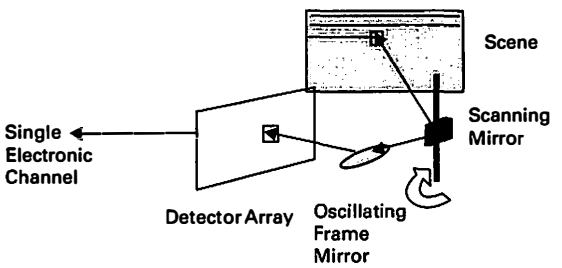


FIGURE 8-24 *Serial scanning*[xxii]

TABLE 8-4 *Advantages and disadvantages of serial scanning*[xxiii]

Advantages	Disadvantages
Small aperture scanner	High scan speed
Minimum number of electronic channels	Large signal bandwidth
Image uniformity	Reliability
Output directly TV compatible	

Parallel Scanning

The AH-1W night targeting system (NTS) FLIR is an example of parallel scanning. It uses a rotary scanner to optomechanically scan the detector array across the scene in a parallel swath; see figure 8-25. One complete rotation of the scanner provides eight fields (or four frames) of thermal imagery.

Each facet provides 160 active lines of thermal imagery for one FLIR field. Every other facet is interlaced to form a 320-line FLIR frame with 2:1 interlace. Adjacent

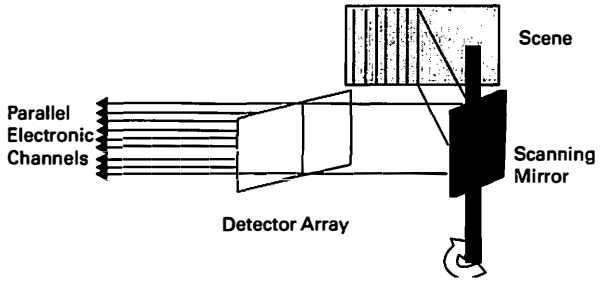


FIGURE 8-25 *Parallel scanning*

active lines are averaged and the average line is inserted between the adjacent active lines to provide 480 total lines of thermal imagery per frame. This is a compatible format for a 525-line TV display.

TABLE 8-5 *Advantages and disadvantages of parallel scanning*[xxiv]

Advantages	Disadvantages
Low scan speed	Large size of detector array with many electronic channels
Small signal bandwidth	Image nonuniformities
	Output not directly TV compatible

Focal Plane Array (Staring)

The UH-1N Star Safire incorporates a 320 × 240 focal plane array to provide parallel processing of the FLIR image. A focal plane array is essentially a checkerboard of sensors that remain stationary with respect to the visual field; see figure 8-26. The visual resolution of any focal plane array system depends upon the ability to shrink the many electronic channels required by 77,000 sensors down to a usable size.

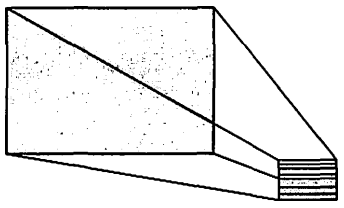


FIGURE 8-26 *Staring array*

Using microchip-etching techniques, each detector in the array is attached to its own data path, thus allowing continuous updating of each individual pixel. This eliminates scanning and virtually all of the moving parts of the system, which generally increases reliability.

It also integrates microscanning (virtual scanning) to perform the interlacing function, thus resulting in a 640 × 480 image for display to the aircrew. Focal plane array systems are the most popular new systems that are being developed. All of the new Generation III FLIR systems use a focal plane array.

TABLE 8-6 *Advantages and disadvantages of staring arrays*

Advantages	Disadvantages
No optical scanning	Large size of detector array
Small signal bandwidth (per channel)	Many parallel electronic channels
	Output not directly TV compatible

Detectors

There are also two general classes of detectors: photon (or quantum) and thermal detectors. Photon detectors convert absorbed photon energy into released electrons. In contrast, thermal detectors absorb energy over a broad band of wavelengths. The energy absorbed by this type of detector causes the temperature of the material to increase. These electrical property variations are measured to determine the amount of incident optical power. Table 8-7 shows the two types of detectors and their different methods described in this chapter.

TABLE 8-7 *EO detectors*

Photon Detectors	Thermal Detectors
Photoconductive	Bolometer
Photovoltaic	Pyroelectric
Photoemissive	

Photon detectors

Photon detectors generate free-charge carriers through the absorption of photons. The absorption occurs without any significant increase in the detector temperature. Detection occurs by a direct interaction of photons with the atomic lattice of the material.

Material parameters that can be changed by this interaction are resistance, inductance, voltage, and current. The concepts presented here apply to detectors that generally sense wavelengths from 0.2μm to 1,000μm. A variety of bandgaps (and corresponding wavelength sensitivities) can be achieved with the selection of materials.

Recall that $E = hv$, where $v = c/\lambda$, h is Planck's constant, and c is the speed of light. The wavelength of a photon has to be short enough that the photon energy exceeds a material's bandgap energy level for absorption.

Intrinsic detectors are those that are constructed from semiconductor crystals with no introduced impurities or lattice defects. Extrinsic detectors have impurities introduced to change the effective bandgap energy.

Cryogenic cooling (with closed-cycle coolers or liquid nitrogen) is typically required for infrared photon detectors to reduce dark-current-generated electron noise. The reduction in temperature reduces the level of dark-current noise to below that of photon-generated signals. Photon-detector responsivity is sometimes characterized by a quantum efficiency that describes the number of electron transitions per absorbed photon.

Intrinsic detectors have quantum efficiencies of around 60%. Typical quantum efficiencies for photoconductors range from 30% to 60%. Photovoltaic and photoemissive detectors have quantum efficiencies of around 60% and 10%, respectively. The detection mechanisms for the three most common types of photon detectors are photoconductive, photovoltaic, and photoemissive detectors.

Photoconductors The photoconductive detector is constructed of semiconductor material that converts absorbed photons into free-charge carriers. The change in free carriers varies the conductivity (and resistance) of the detector. A simple circuit provides a bias current that reflects the change in conductivity.

Photoconductors sense optical power with a change in conductance (or inverse resistance). Longer wavelengths can be detected by doping a pure material with impurities, giving an extrinsic detector. A P-type extrinsic detector decreases the conduction band energy level, and an N-type extrinsic detector raises the valence band energy level. The result is a smaller bandgap material that can detect longer wavelengths.

Quantum efficiencies are lower (around 30%) for extrinsic detectors because the dopant is less abundant than the host material. For photoconductors, an increase in the incident optical power increases the number of free-charge carriers, so the effective resistance of the detector decreases. The differential charge, dq (in Coulombs), provided by the impinging photons is given by

$$dq = E_p \eta_q w t_q e l_d \tag{8-12}$$

where:

E_p = incident irradiance in photons per second per square centimeter
w = detector width in centimeters
l_d = detector length in centimeters
η_q = quantum efficiency
t_q = charge carrier mean lifetime in seconds
e = charge on an electron (1.602×10^{-19} Coulombs)

Photovoltaic The photovoltaic detector absorbs photons at a P-N junction. Hole-electron pairs are generated that alter the junction voltage. The change in voltage describes the amount of optical power on the detector. The photovoltaic detector does not require any external bias voltage. The inherent voltage of the P-N junction changes without bias and can be read out directly.

Like photoconductors, the photovoltaic detector has spectral characteristics that are derived from the material's molecular energy levels. Photovoltaic detectors are constructed with the formation of a P-N type junction in a semiconductor. The P-N junction has an inherently generated voltage and does not require a bias voltage or current. A typical photovoltaic detector energy diagram is shown in figure 8-27.

The photons are absorbed at the junction of the P- and N-type materials, an area called the depletion region. The junction is the same as a diode with the added effect that voltage is generated with illumination.

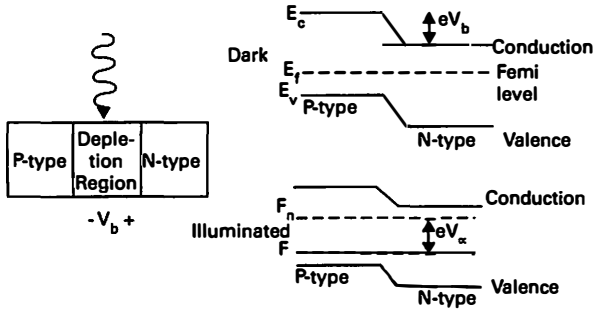


FIGURE 8-27 *Photovoltaic energy diagram*[xxv]

The energy diagrams of figure 8-27 show the diode in the dark and illuminated states. The voltage in the dark state is related to the differences in conduction (and corresponding valence) energy levels for the P- and N-type materials across the depletion region.

The electrical characteristics of photovoltaic detectors are similar to those of a diode. Because of this similarity, photovoltaic detectors are sometimes referred to as photodiodes. Photovoltaic detectors offer several advantages over photoconductor detectors: more responsiveness, simpler biasing, and a better theoretical signal-to-noise ratio (SNR). The photovoltaic detector is normally operated in the open-circuit mode where the photo-induced voltage is measured directly. This mode has SNR advantages over a system that requires external amplification.

Photoemissive Photoemissive devices are different from photoconductors and photovoltaic detectors in that an electron physically leaves the detection material when a photon is absorbed. Usually placed in a vacuum tube with high voltage, a photocathode absorbs the photon, and the energy absorbed is enough (higher than the electron work function) to release the electron in the vacuum chamber. The voltage potential moves it toward the anode. Once electrons reach the anode, they are measured as a current that reflects the optical flux on the photocathode.

Photoemissive detectors release electrons from the photosensitive surface during the detection process. The incident photon of energy, hv, provides the energy necessary to release an electron from the surface. The process is shown in figure 8-28 and is described as follows:

The photon enters a glass envelope that houses the photocathode and anode. Usually the envelope comprises a vacuum tube where conductive material deposited on the glass surface acts as the cathode. The photon energy is absorbed, producing a free electron that is accelerated to the anode through the vacuum. Only electrons that have sufficient energy to overcome the work function of the photocathode material result in released electrons. The electrons are collected at the anode, and the corresponding

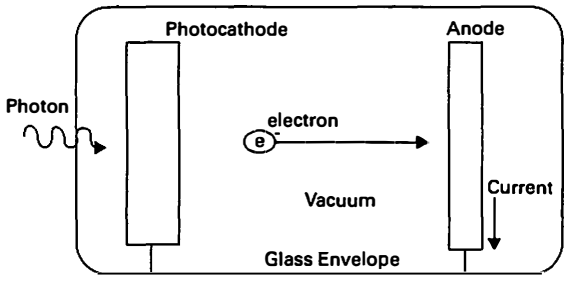


FIGURE 8-28 *Photoemissive detection*[xxvi]

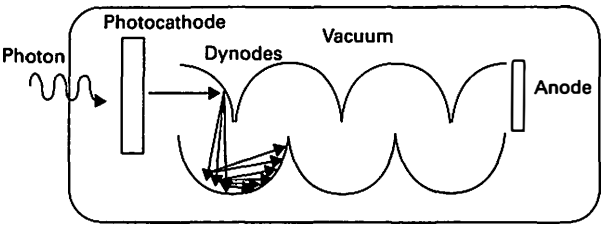


FIGURE 8-29 *Photomultiplier*[xxvii]

current is measured to determine the amount of light on the photocathode.

Photocathodes are designed with high optical absorption and small work functions. The photomultiplier tube (PMT) is a special case of the photoemissive detector. The PMT is used when large gains are required to sense very low light levels. The PMT consists of a photocathode, electron multipliers (dynodes), a collection anode, and an evacuated glass or metal envelope. The photomultiplier is shown in figure 8-29.

An electron is released from the photocathode and is accelerated to the first dynode. The electron strikes the first dynode with enough energy to release a number of electrons that are accelerated to the next dynode. This multiplication occurs through a number of dynodes, resulting in a large number of electrons collected on the anode. The gain of a PMT is estimated by:

$$G = \chi^n \text{ (unitless)} \qquad (8\text{-}13)$$

where:

G = current gain
χ = secondary emission ratio for dynodes
n = number of dynodes

A nine-stage PMT with a secondary emission ratio of 8 gives a gain of around 10^8. This large gain and low noise make the PMT an attractive detector for many applications.

Thermal Detectors

Thermal detector materials have at least one inherent electrical property that changes with temperature. This temperature-related property is measured electrically to determine the incident radiation on the detector. These detectors typically operate at room (ambient) temperature and do not require cryogenic cooling. The change in temperature produces one of three electrical property changes in the detector material: resistance, voltage, or capacitance. The electrical property change groups thermal detectors.

Thermal detectors include bolometers and pyroelectric detectors. Bolometers are sometimes called thermistors, because the material provides a resistance that changes

with temperature. Semiconductors are used with high-temperature coefficients of resistance. The material is usually a long, narrow strip of material with a highly absorbent coating. The electrical resistance of the strip is measured with a bias current to determine the amount of absorbed radiation.

In the past, thermal detectors have been characterized by their slow response time. Significant developments in recent years, however, have made these detectors viable candidates for many applications. Thermal detectors today exhibit many good characteristics, including quick response time, low noise, and linear operation. Even though the sensitivity of thermal detectors has improved dramatically in recent years, they are still behind the performance of photon detectors.

Bolometers Bolometer detection is derived from a change in the resistance of the detector material. A simple bolometer detector consists of a thin, blackened metal or semiconductor filament, as shown in figure 8-30.

The change in the bolometer resistance due to incident radiation is given by:

$$\Delta R_b = \Delta R_0 \alpha \Delta T \qquad (8\text{-}14)$$

where:

ΔR_b = change in bolometer resistance (ohms)
ΔT = filament's temperature change (K)
R_0 = strip resistance when $\Delta T = 0$ (ohms)
α = temperature coefficient of resistance ($^1/_K$)

The temperature coefficient is positive for metals and negative for semiconductors. In the absence of incident radiation, there are changes in the detector from conduction from the surrounding environment. The changes are statistical in nature and contribute to detector noise. For this reason, the detector is isolated from the local environment to help ensure that signals are a result only of incident radiation. Isolation is achieved through special manufacturing techniques and by placing the bolometric detector in a vacuum.

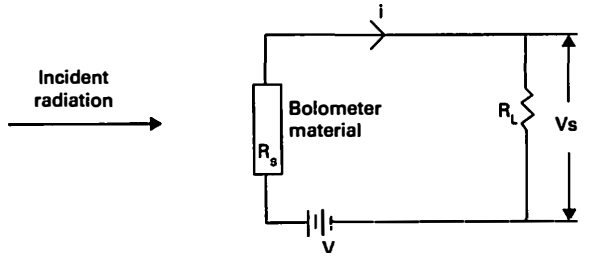


FIGURE 8-30 *Bolometer configuration*[xxvii]

Pyroelectric Pyroelectric detectors are the most common thermal detectors. The pyroelectric detector is a ferroelectric material between two electrodes. The material exhibits a change in polarization with temperature. It can be thought of as a capacitor whose capacitance is temperature sensitive.

As with any capacitor, measurement requires an AC signal, so most pyroelectric detectors are chopped to vary the incident flux. Pyroelectric detectors have a change in surface charge with a change in temperature. This pyroelectric effect, as it is called, produces a current that is proportional to the change in temperature:

$$i_p = p_T A_a \frac{dT}{dt} \qquad (8\text{-}15)$$

where:

i_p = electric current (amps)
p_T = pyroelectric coefficient at the temperature T Kelvin ($C/cm^2 \cdot K$)
A_d = surface area of the detector (cm^2)
dT/dt = time rate of temperature change (K/sec)

When a constant radiation is applied to the detector, no current can be sensed. A chopper must be used with the detector in order to provide changing temperatures that produce a current to sense a constant signal. Pyroelectric detectors are in essence capacitors. The responsivity for a pyroeletric detector is similar to that of the bolometer.

Pyroelectric materials are ferroelectric crystals that show spontaneous changes in polarization along one axis. Two materials used for this kind of detector are PZT (doped lead zirconate titanate) and BST (barium strontium titanate). The polarization of these materials varies with the temperature and is characterized by the pyroelectric coefficient. This family of thermal detectors is characterized by fast response times across a broad spectrum.

Charge-Coupled Devices

The charge-coupled device (CCD) has become very familiar due to the popularity of handheld video cameras operating in the visible band, although there are also infrared CCDs (IRCCD). The term *CCD* does not refer to a specific detector type but rather to a detector readout architecture. The basic concept of charge-coupling is that once the photo-generated charge is collected in potential wells, it is selectively transferred or read out of each detector location in a serial process.

CCDs are typically metal-oxide-structure (MOS) devices. The basic structure for these gates is shown in figure 8-31. When a positive voltage is applied—V_g

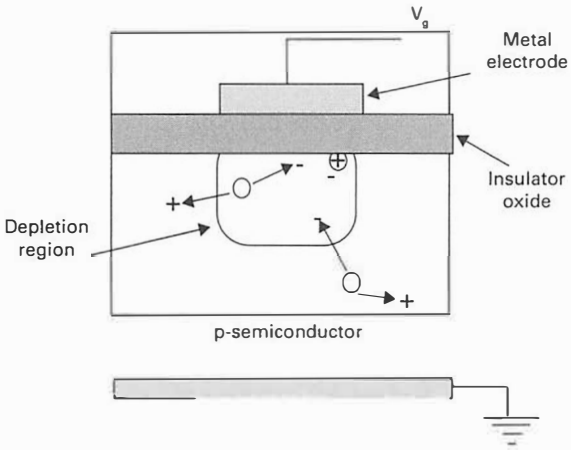


FIGURE 8-31 *Metal oxide structure*[xxix]

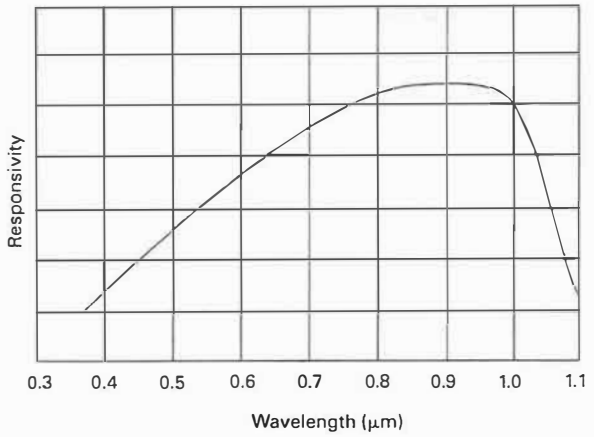


FIGURE 8-32 *Responsivity curve*[xxx]

between the top metal electrode and the substrate—the majority carriers (holes) are moved from the oxide layer, creating a depletion region at the interface. If a photon is absorbed into the region, then an electron-hole pair is produced. For the photon to be absorbed, its energy must be greater than that of the bandgap.

The CCD is an analog shift register in which an entire sequence of charge packets can be shifted simultaneously along the surface of the chip by applying clock voltages (V_g). The electric charge varies from one packet to the next. The sequence of charge packets are scanned as they move past an output electrode connected to an on-chip amplifier. To read out the charge packets, each component in the array is connected to sequential shift registers on the periphery. A clock that controls the time delay then clocks the packets out of the shift registers.

Detector Responsivity

The fundamental purpose of a detector is to convert an optical signal into an electrical current or voltage. The term *responsivity* is used to describe the amplitude of the electrical signal with respect to the incident flux. Responsivity is a function of wavelength and temporal frequency and can be determined by:

$$R(\lambda,f)=\frac{V_{out}}{\Phi_{incident}}=\frac{V_{out}}{E(\lambda,f)A_d} \tag{8-16}$$

where:

$R(\lambda, f)$ = responsivity (volts/watts)
V_{out} = detector output voltage
$\Phi_{incident}$ = incident flux (watts)
$E(\lambda, f)$ = incident irradiance (watts/cm^2)
A_d = detector area (cm^2)

Current responsivity would have units of Amperes per watt. A typical photon detector responsivity curve is shown in figure 8-32.

The responsivity is typically measured by the detector manufacturer and provided as a specification or calibration document. The band-averaged responsivity over a spectral band $[\lambda_1, \lambda_2]$ is given by:

$$R=\frac{\int_{\lambda_1}^{\lambda_2}\Phi(\lambda)R(\lambda)d\lambda}{\int_{\lambda_1}^{\lambda_2}\Phi(\lambda)d\lambda} \tag{8-17}$$

where:

R = responsivity (volts/watts)

This value is used frequently in cases where the incident flux and the responsivity are not expected to vary significantly over the spectral band. For EO systems, the responsivity is given frequently in terms of photometric units:

$$R_{photopic}=\frac{\int_{\lambda_1}^{\lambda_2}\Phi(\lambda)R(\lambda)d\lambda}{683\int_{\lambda_1}^{\lambda_2}V(\lambda)\Phi(\lambda)d\lambda} \tag{8-18}$$

where:

$R_{photopic}$ = responsivity (volts/Lumens)
$V(\lambda)$ is the human eye's photopic efficiency

Note: "683" represents the peak visual efficiency (Lumens/watts).

The use of equation 8-18, photometric responsivity, is valid for EO systems that compare similarly in response to the human eye. However, as the system response moves farther from the human eye reference large differences can occur between actual and predicted values.

Sensor Bandwidth

The selection of the sensor bandwidth should be such that the best sensitivity is achieved throughout the expected range of operating conditions and target signatures. Atmospheric windows usually limit the selection within the 0.5μm to 2.5μm, 3μm to 5μm, or 8μm to11μm bands.[xxxi] The sensor bandwidth will determine the amount of energy incident upon the aperture or FOV, which the sensor will detect.

When the upper and lower wavelengths of the sensor bandwidth are multiplied by the absolute target temperature and are applied to the universal blackbody curve (figure 8-33), the percentage of total incident energy may be determined. This is known as the bandwidth factor.

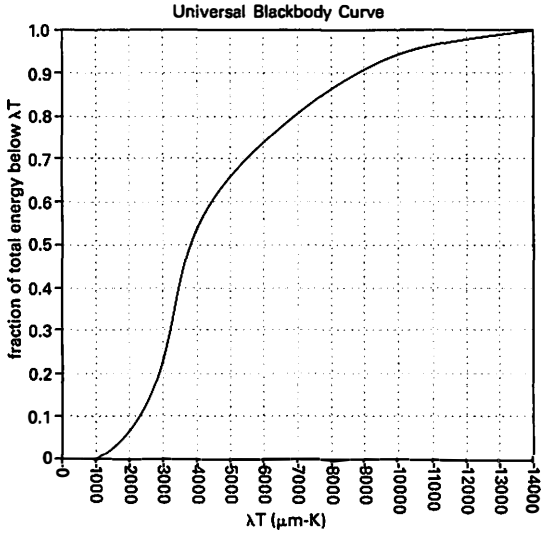


FIGURE 8-33 *Universal blackbody curve*

EXAMPLE 8-3: For a detector band width from 8μm to 14μm, what fraction of the energy radiated from an object at 500°K can be processed by the receiver due to bandwidth limitations? (That is, calculate the bandwidth factor.)

SOLUTION: First obtain λT for the upper and lower limits of the bandwidth:

$$\lambda_u T = 14\mu m \times 500°K = 7{,}000\mu m°K$$

$$\lambda_l T = 8\mu m \times 500°K = 4{,}000\mu m°K$$

Entering figure 8-33 at a $\lambda_u T$ of 7,000μm°K and reading the fraction of radiated energy axis $f_u = 0.81$ implies that 80% of exitance is radiated below 7,000 μm°K.

The lower bandwidth limit $f_l = 0.51$ again implies that 51% of exitance is radiated below at $\lambda_l T$ of 4,000μK.

Therefore, the bandwidth factor:

$$F = f_u - f_l$$

$$F = 0.81 - 0.51 = 0.3$$

Following the definition of the bandwidth factor, this states that 30% of the radiated power per unit area, exitance, is radiated in the bandwidth of this particular receiver.[xxxii]

Our eyes operate similarly. For example, for objects around 30°C, our eyes are able to "see" less than 10% of the incident radiation.

Sensor Response

Sensor response is the efficiency of the conversion of the received energy to an electrical signal. The sensor output is commonly measured in volts, but the input radiation can be expressed in a variety of forms. One of the most common expressions of infrared incident radiation is temperature differential. The response is then the voltage-out produced by the differential.

The responsivity of a laser detector is specified as quantum efficiency—the percent of photon hits that result in the dislodging of an electron. When a photon strikes laser-sensitive material, an electron may be freed to contribute to an electrical current.[xxxiii]

Sensitivity

The sensitivity of a sensor is determined by a large number of factors including optical transmission, bandwidth, aperture size, temperature differentials, and sensor response. These characteristics will dictate the minimum discernable signal (MDS), or $S_{min,}$ necessary for the sensor to detect the target. These terms are identical to what was used in chapters 3 and 4 when describing radar performance.

It is important to remember that as the sensitivity increases, the MDS and S_{min} decrease. Sensor sensitivity is the only factor in determining maximum detector range over which a system designer has control. A target's incident radiation must produce sufficient signal-to-noise ratio for detection.

Target Detection

To determine the exact amount of radiation originating from an object and propagating to the sensor detector (i.e., no path radiance and background radiance), one must consider the atmospheric transmission as a function

of wavelength. The radiation received by the detector within a sensor can be expressed by:

$$P_{sensor} = \int_{\lambda_l}^{\lambda_u} \frac{M_{source}(\lambda)}{\pi} A_{source} \Omega_{sensor} \tau_{atm}(\lambda) \tau_{sensor}(\lambda) d\lambda \quad (8\text{-}19)$$

where:

$\frac{M_{source}(\lambda)}{\pi}$ = source (Lambertian) radiance in (w/cm²/sr/μm)

A_{source} = source area seen within the sensor field of view in (cm²)

Ω_{sensor} = solid angle of the sensor entrance pupil as seen by the source in (sr)

$\tau_{atm}(\lambda)$ = transmission of the atmosphere between the source and the sensor (unitless)

$\tau_{sensor}(\lambda)$ = transmission of the sensor optical path (unitless)

The question, "What targets can I see with IR viewers or low-light-level (LLL) systems?" is like the question, "How far can I see?" The answer to each is "It all depends. . . ." Still, it is possible to indicate a few general order-of-magnitude statements.

In all systems involving transmission or reception of electromagnetic radiation, the undesired factor of noise must be considered. Noise sources internal to IR systems are thermal fluctuations in the molecular structure of the detecting surface, or in the electronic circuitry associated with the system.

A detector's noise-equivalent-power (NEP) is the input radiant flux in watts necessary to produce an output signal equal to the detector noise. A detector with lower NEP has higher sensitivity.

The source will contribute power to the detector as a function of the source temperature, and the detector will radiate away power in accordance with its own temperature. Another important factor is the difference between the source temperature and the background temperature, which is termed background noise.

Thermal detection, like all other detection schemes, is a differential process. The actual detection problem is more complicated than suggested here. More than the source temperature, emissivity, and reflectivity must be considered. The thermal level and gradients in the surrounding environment need to be considered as well.

As the temperature of the source approaches that of the detector, self-noise can mask the target. This makes cooling the detector important in order to lower self-noise. Recall that Wien's Displacement Law relates the temperature and wavelength of maximum radiation.

A system that cannot detect a target because the temperature differential between the source and the detector is less than the detector's required minimum detectable signal is said to be "background limited." A system that cannot detect a target because the signal radiated by the target (source) is masked by the self-noise of the detector is said to be "noise limited."[xxxiv]

Simplified IR Range Equation

Combining the principles discussed so far, it is possible to make an estimate of maximum detection range for a target. Using logic similar to that in the radar range equation, the IR equation is developed as follows. (Only a one-way trip needs to be considered however.)

Recall that the exitance of a target is

$$M = \varepsilon\sigma\left(T^4 - T_e^4\right) \quad (8\text{-}20)$$

where:

T = absolute temperature of the source (K)

T_e = absolute temperature of the environment (K)

ε = emissivity (unitless)

σ = Stefan-Boltzmann constant (w/m²K⁴)

From equation 7-2 the total flux Φ is:

$$\Phi_{TOT} = \varepsilon\sigma\left(T^4 - T_e^4\right)A_{tgt} = MA_{tgt} \quad (8\text{-}21)$$

Assuming a point source and spherical spreading, the incidence at the detector at some range, R, is:

$$E = \frac{\Phi_{TOT}}{4\pi R^2} = \frac{\varepsilon\sigma\left(T^4 - T_e^4\right)A_{tgt}}{4\pi R^2} \quad (8\text{-}22)$$

Multiplying by the effective area of the detector (A_e):

$$E = \frac{\Phi_{TOT}}{4\pi R^2} = \frac{\varepsilon\sigma\left(T^4 - T_e^4\right)A_{tgt}A_e}{4\pi R^2} \quad (8\text{-}23)$$

However, because the detector "sees" only a portion of the incident energy, the bandwidth factor, F, must be applied.

$$E_{actual} = \frac{F\varepsilon\sigma\left(T^4 - T_e^4\right)A_{tgt}A_e}{4\pi R^2} \quad (8\text{-}24)$$

Setting the incident energy detected equal to S_{min} and then solving for the range, R, the following is obtained:

$$S_{min} = \frac{F\varepsilon\sigma\left(T^4 - T_e^4\right)A_{tgt}A_e}{4\pi R^2} \quad (8\text{-}25)$$

$$R = \sqrt{\frac{F\varepsilon\sigma\left(T^4 - T_e^4\right)A_{tgt}A_e}{4\pi S_{min}}} \quad (8\text{-}26)$$

EXAMPLE 8-4: A NASA substation has been directed to track a space vehicle during reentry. Given the following data, at what range will the target be detected?

Surface area of vehicle: 125 m^2
Detector bandwidth: 3μm to 5μm
Heat-shield emissivity: 0.7
Vehicle skin temperature: 575°C
T_e at reentry altitude: −180°C
IR detector aperture: 1.5 m^2
IR detector S_{min}: 61.1×10^{-7} w
$\sigma = 5.67 \times 10^{-8}$ w/m^2K^4

SOLUTION: The spacecraft exitance is:

$$M = \varepsilon\sigma\left(T^4 - T_e^4\right)$$

Converting to absolute temperature:

$$T_{skin} = 575°C + 273° = 848°K$$
$$T_e = -180°C + 273° = 93°K$$
$$M = (0.7)(5.67 \times 10^{-8} s/m^2K^4)((848°K)^4 - (93°K)^4$$
$$M = 20{,}500 \text{ watts/m}^2$$

The total power radiated:

$$\Phi_{TOT} = M \times \text{area}_{vehicle}$$
$$= 2.0521 \text{ w/m}^2 \times 125 m^2$$
$$= 2.57 \times 10^6 \text{ watts}$$

The detector can use only that energy within its bandwidth of 3–5μm.

For band λ_u to λ_l of 5μm to 3μm,

$$\lambda_u T = 5\mu m \times 848°K = 4{,}240\ \mu m°K$$
$$\lambda_l T = 3\mu m \times 848°K = 2{,}544\ \mu m°K$$

Using the universal blackbody curve, the determination is that 0.55 of the total energy is below 4,240μm°K and that 0.13 of the total energy is below 2,544μm°K. This information is used to calculate the bandwidth factor, F.

$$F = f_u - f_l$$
$$F = 0.55 - 0.13 = 0.43$$

The maximum range using equation 8-26 is:

$$R = \sqrt{\frac{0.43(2.565 \times 10^6 w)(1.5 \times m^2)}{(4\pi)(61.1 \times 10^{-7} w)}}$$
$$= 146.8 \text{ km}$$

Night Vision Devices (NVDs)

There are two ways of improving night vision. The first is to increase the amount of light reaching the eye, as with a telescope or flashlight. The second is through light amplification, by creating a visible screen image from normally imperceptible radiation, as in a night vision imaging system (NVIS) or a forward-looking infrared (FLIR) camera.

Night vision scopes and binoculars "amplify" existing light, allowing the human eye to see in conditions otherwise too dark for the naked eye. In the night vision system, available light (photons) is collected by the objective lens and focused on the functional heart of the device, the image intensifier. It is the conversion of photon to electrons in the image intensifier that leads to the display of an enhanced image.

Night vision systems improve the ability to see in low-light conditions, but they cannot provide a useful image in complete darkness because there is no available light to amplify. In caves, warehouses, or any other location where no ambient light is available, an accessory infrared illuminator may be added to the night vision system. The illuminator provides a source for the system to amplify. Since the light emitted by the illuminator is in the near-infrared range, it is invisible to humans; see figure 8-34.

An infrared illuminator can provide a great deal of utility, even in conditions where only a small amount of light is available. It can provide enhanced imaging capability in very-low-light conditions, and offer more consistent performance as the user moves from low-light to very-low-light locations.

Image Intensifier (I2) Tube

An image intensifier (I2) is an electro-optical device used to detect and intensify energy in the visible and near-infrared regions of the electromagnetic spectrum. Referring to

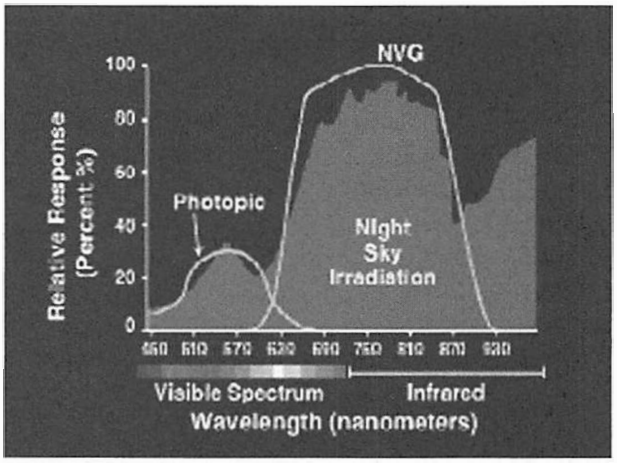


FIGURE 8-34 *Spectral response of the eye (photopic) and NVGs compared to night sky irradiaton*

figure 8-35, notice that reflected energy from the terrain enters the night vision device (NVD) and is focused by the objective lens onto a photocathode, which is sensitive to portions of both visible and near-IR energy. The photocathode then releases electrons proportionate to the number of photons incident upon the photocathode.

The electrical field produced by the NVG power supply accelerates the electrons away from the photocathode and toward a phosphor screen, which emits light proportional to both the number and the velocity of electrons that strike it. Thus, the electrical field applied between the photocathode and the phosphor screen accelerates the electrons and results in an increase in brightness of the NVG image.[xxxv]

In the night vision lexicon, "generation" designates the major advancements in technology. The higher the generation, the more sophisticated the night vision technology. The following is a synopsis of the development of this technology.

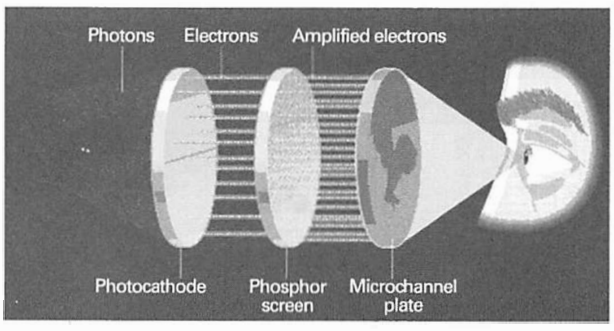


FIGURE 8-35 *Typical stages of night vision image intensification.*[xxxvi]

Generation 0 Image Converters

These devices were first experimentally used in World War II and the Korean War and consisted of little more than large IR searchlights with infrared detectors. Although they were not image intensifiers like current NVDs, their development spawned the drive to develop passive night imaging systems. Gen 0 devices were bulky, fragile, power hungry, short ranged, and operated as active IR sensors.

Models such as the M1 Sniperscope, used on Okinawa, and the M3 Infrared scope used in Korea—(see figure 8-36)—did little to intensify the amount of light reflecting off a target. They simply converted the infrared energy into visible images that were usable by the infantry soldier.



FIGURE 8-36 *Gen 0 M3 Infrared Sniperscope with .30 caliber carbine*

These devices required the illumination of the target area by an active infrared light source mounted on or near the sensing device. The night vision scope collected the infrared light reflected off the target, which generated an image for the user, often a sniper. Unfortunately, the light source was visible to anyone with a night vision device, highlighting the position of the unit. This soon became a liability, as neither side wanted a large "IR spotlight" marking its position.

First-Generation Image Intensifiers

Generation I night vision devices permitted viewing at night without the aid of any light source other than the dim glow of the moon, stars, or even faint sky. Completely passive night vision technology was a giant leap ahead and greatly enhanced night tactics. The technological improvements included vacuum-tight, fused fiber optics for good center resolution and improved gain; multialkali photocathodes; and fiber-optic input and output windows.

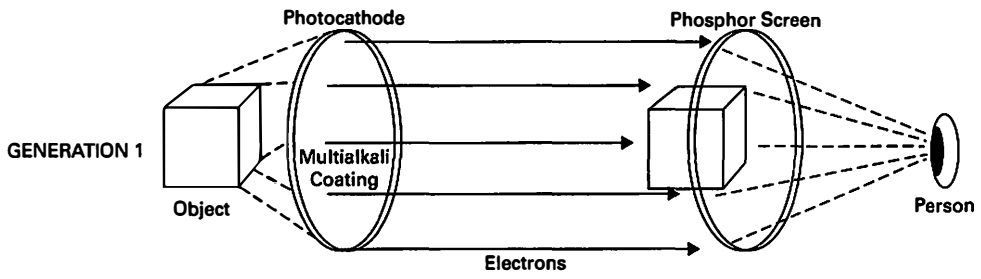


FIGURE 8-37 *Gen I (I2) electrostatic tube technology*

Gen I intensifiers were composed of simple electrostatic tubes. Light reflected from the subject struck a multialkali-coated photocathode, which produced electrons where the light struck it. These electrons were displayed on the phosphor screen.

Even though they were a marked improvement over Gen 0, Gen I tubes lacked the sensitivity and light amplification necessary to see below full moonlight.

The first operational use of Gen I technology was the starlight scope used by snipers in Vietnam. Fielded in the late 1960's, these systems amplified light through use of a three-stage configuration of the simple image intensification tube; see figures 8-37 and 8-38. Three image intensifier tubes connected in series amplified the ambient light about 1,000 times.

Gen I equipment was large, inefficient, and cumbersome. It also suffered from poor reliability. Additionally, it was susceptible to "blooming." Blooming is a tendency for the tube to "shut down" if a bright light source appears in the device's field of view. This causes the overall contrast of the intensified image to be reduced.

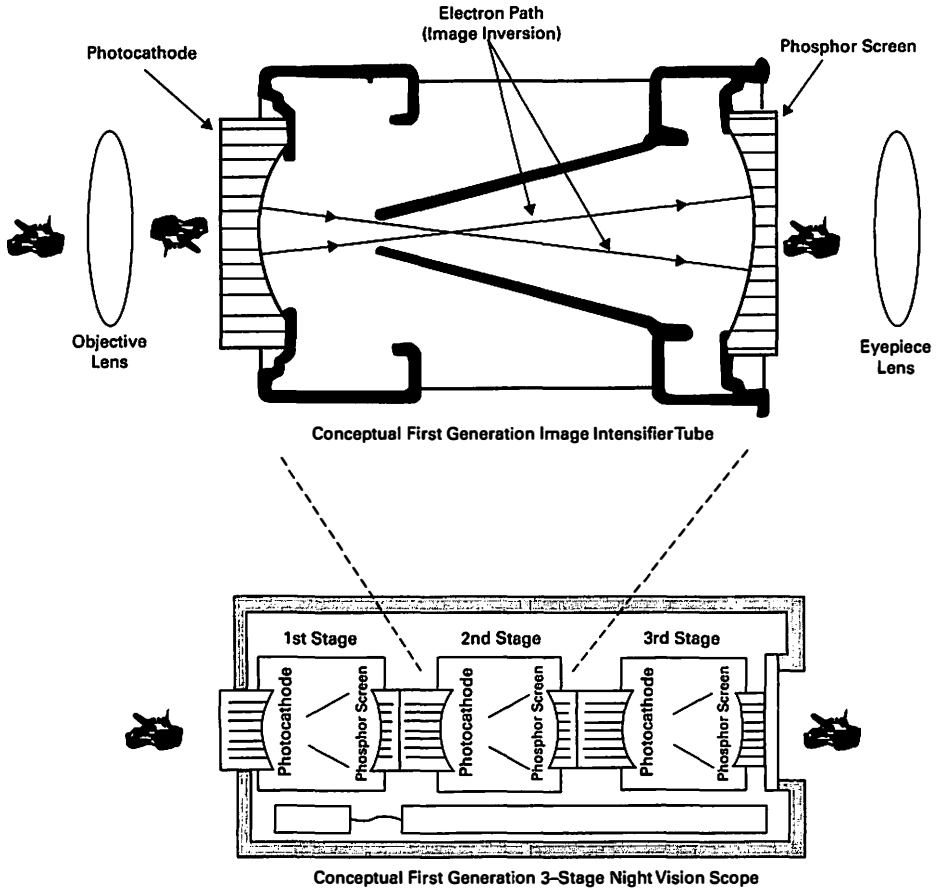


FIGURE 8-38 *GEN I (I2) tube technology*[xxxvii]

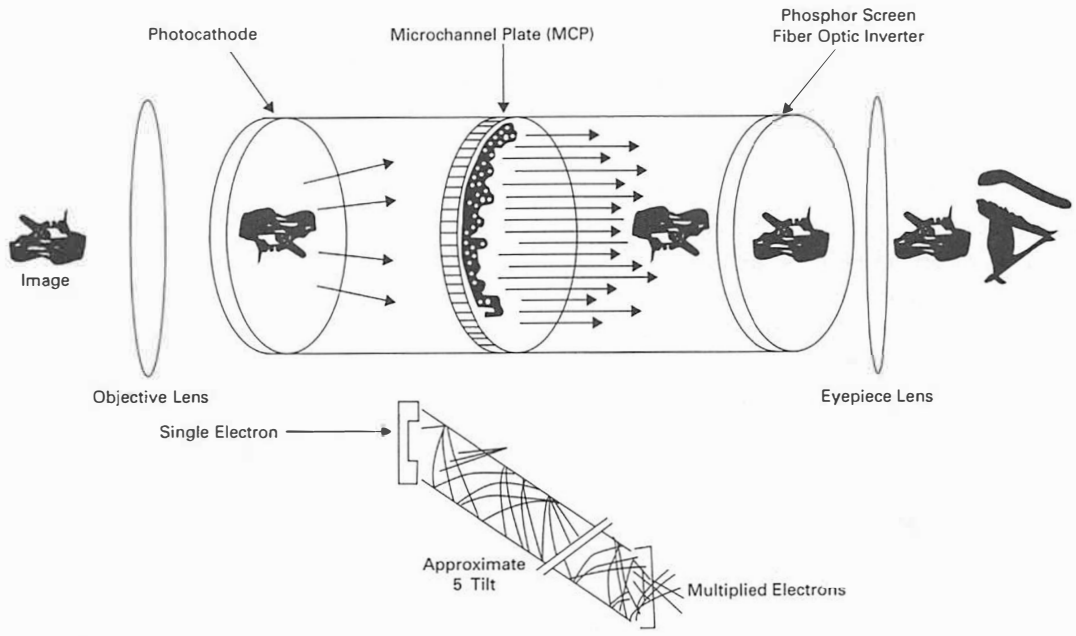


FIGURE 8-39 *Gen II tube technology*[xxxix]

Direct viewing of light could damage the image tube and cause spotting or "scars" on the tube face. Other problems were streaking and distortion. During the Vietnam War, Gen I systems were utilized for a variety of missions including observation, night "sniping," and nighttime command and control.

Second-Generation Image Intensifiers

In the late 1960s and early 1970s the development of the microchannel plate (MCP) led to major improvements in image intensifier tubes and the eventual birth of Gen II devices. The biggest gain in Gen II devices was their ability to see in extremely low-light conditions, such as 1/4 moonlight nights. This increased sensitivity was directly due to the addition of the MCP electron multiplier to the image-intensifier tube.

The MCP is a thin, lead-silicate glass wafer composed of thousands of individual hollow glass fibers that have been fused into a single bundle and fabricated into a flat, cylindrical plate. The interiors of the hollow glass fibers are coated with a conductive layer and silica surface to boost the acceleration of the electrons emitted from the cathode; see figure 8-39.

Because of the development of the MCP the size and weight of NVD was greatly reduced. Much of the size invested in the first generation tube was necessary due to the voltage required for sufficient electron velocity to produce the desired gain; this need was reduced with the use of the energy efficient MPC. Additionally, the amplification provided by the MCP eliminated the need for back-to-back (staged or cascaded) tubes, which also decreased the size and improved the image quality.[xxxviii]

However, the largest improvement that the MCP caused was the increase in gain of the intensifier tubes from 1,000 in Gen I to approximately 20,000 in Gen II. Since the MCP actually increased the number of electrons instead of just accelerating the original ones, the images were significantly less distorted and brighter than earlier-generation NVDs.

Reduction of the size of the image intensification tube allowed Gen II devices to be the first tubes small and light enough to be placed into systems that could be head mounted, such as the AN/PVS-5 shown in figure 8-40.

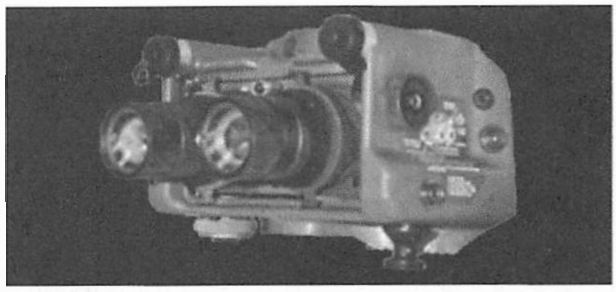


FIGURE 8-40 *Gen II AN/PVS-5*

Although second-generation tubes were still somewhat susceptible to "blooming," the MCP significantly reduced this tendency. The MCP confines the saturation to individual channels, thus localizing contrast degradation within the FOV. This degradation appears as a "halo" around a bright source.[xli]

Gen II systems also incorporated two additional improvements: automatic brightness control and bright source protection. These functions automatically adjusted the gain or amount of light coming from the phosphor screen to the user's eye. Additionally, Gen II devices had better image enhancement, improved contrast and resolution, and longer tube life.

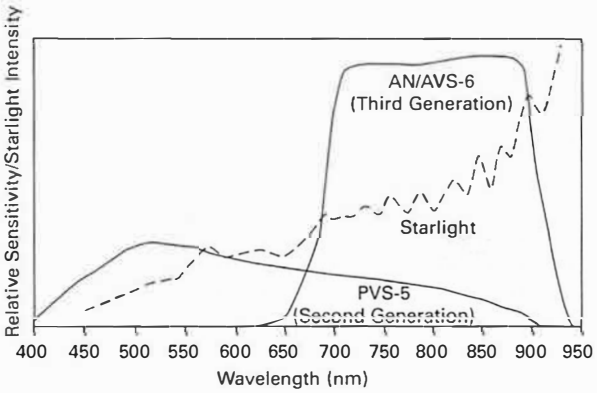


FIGURE 8-41 *Spectral sensitivity of second- and third-generation image intensifier tubes*[xl]

Third-Generation Image Intensifiers

In Gen III, the multialkali photocathode (S-20) was replaced by a gallium arsenide (GaAs) photocathode, and an aluminum-oxide ion-barrier film was applied to the MCP. This led to significant performance enhancements. Figure 8-41 shows that the GaAs photocathode is uniquely sensitive beyond 800 nanometers, the near-infrared region where both night sky illumination levels and environmental contrast ratios are the greatest.

Since the spectral response of the GaAs-based photocathode matches the available natural illumination in the night sky, Gen III tubes truly provided the first "starlight" capable systems.[xlii] The Gen III intensifier's gain of 35dB to 47dB, coupled with the spectral response shift to the red region, resulted in greatly improved signal-to-noise ratios. This resulted in a threefold improvement in visual acuity and detection distances. The advertised image acuity using only a quarter moon is 20/40.

Additional improvements in Gen III tubes included the addition of a fiber-optic bundle between the microchannel plate and the phosphor screen; see figure 8-43. This fiber-optic bundle directly channeled the electrons coming through the MCP plate to the phosphor screen, resulting in a more linear electron flow for higher resolution and

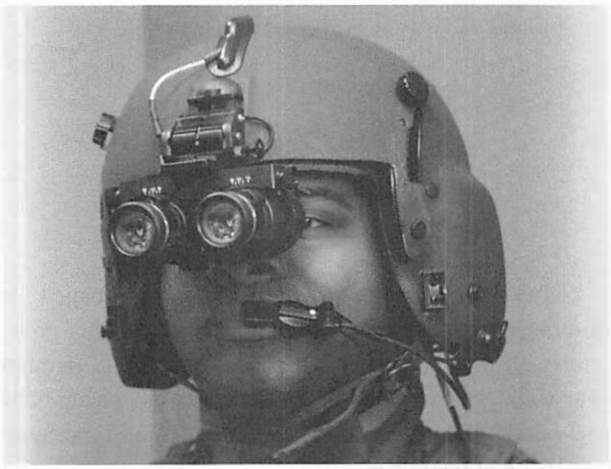


FIGURE 8-42 *Gen III ANVIS-6*

light amplification. Finally, the ion-barrier film increased the operational life of the tube from 2,000 hours (Gen II) to 10,000 hours (Gen III).

Fourth-Generation Intensifiers

Gen IV image intensifier tubes are beginning to appear; see figure 8-45. They eliminate the ion barrier used with

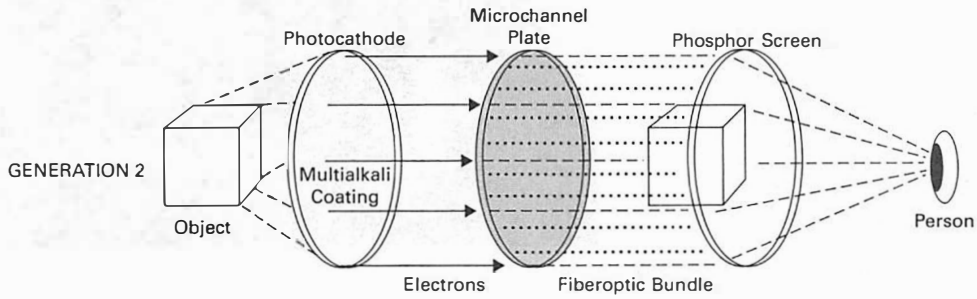


FIGURE 8-43 *Gen III tube technology*

	Full Moon .1 lux	Half Moon .05 lux	Quarter Moon .01 lux	Starlight .001 lux	Overcast .0001 lux
G3	890 yds	890 yds	850 yds	580 yds	220 yds
G2	690 yds	690 yds	650 yds	430 yds	160 yds
Without Night Vision	250 yds	150 yds	50 yds	*	*

*not measurable

FIGURE 8-44 *Estimated detection ranges for a 6 ft person*

Gen III tubes, which greatly simplifies the manufacturing, reducing costs. They also offer an improved signal-to-noise ratio and better performance in high-light-level situations.

Gen IV tube technology will further reduce the amount of blooming in high-light conditions. The first application of this technology will be a day/night sniperscope for the U.S. Special Forces.

Summary

Imaging systems are an important sensor for numerous weapons systems. Most sources or targets emit over a broad spectrum of wavelengths and are termed broadband or thermal emitters. There will be a peak wavelength for any target and sensors will be optimized for that wavelength. For example, an air-to-air missile guiding on a hot jet engine would be more focused on the 3–5μm band, and an air-to-ground missile guiding a tank would be more focused on the 8–12μm band, which is at a much cooler temperature. Other sources, called selective radiators, emit energy at a single or narrow band of wavelengths. Examples of selective radiators are exhaust gases and lasers.

Target emissions are always a combination of reflected and emitted energy. However, depending on the ambient conditions, sensor bandwidth, and the target itself, certain portions of the spectrum may be much more apparent than others. Some targets may be nearly invisible to the naked eye but readily apparent to a sensor in a different sensitivity band. Due to absorption, some wavelengths will be more visible than others.

There are two basic categories of sensors, light amplification and thermal sensors. Light amplification sensors focus on the visible and NIR bands. They require some minimal amount of reflected light to operate. Thermal sensors focus solely on the thermal energy emitted by the target.

Most sensors are in arrays of smaller sensors, something akin to a fly's eye. Each smaller sensor has a field of view, or range of viewing angles, called an instantaneous field of view, or IFOV. The combination of small sensors forms an overall sensor field of view referred to as the FOV. *Depth of focus* is another important optical term, and it relates to the focal length of the lens and its field stop. Although these are usually thought of as visible light properties, these same properties apply throughout the EO spectrum.

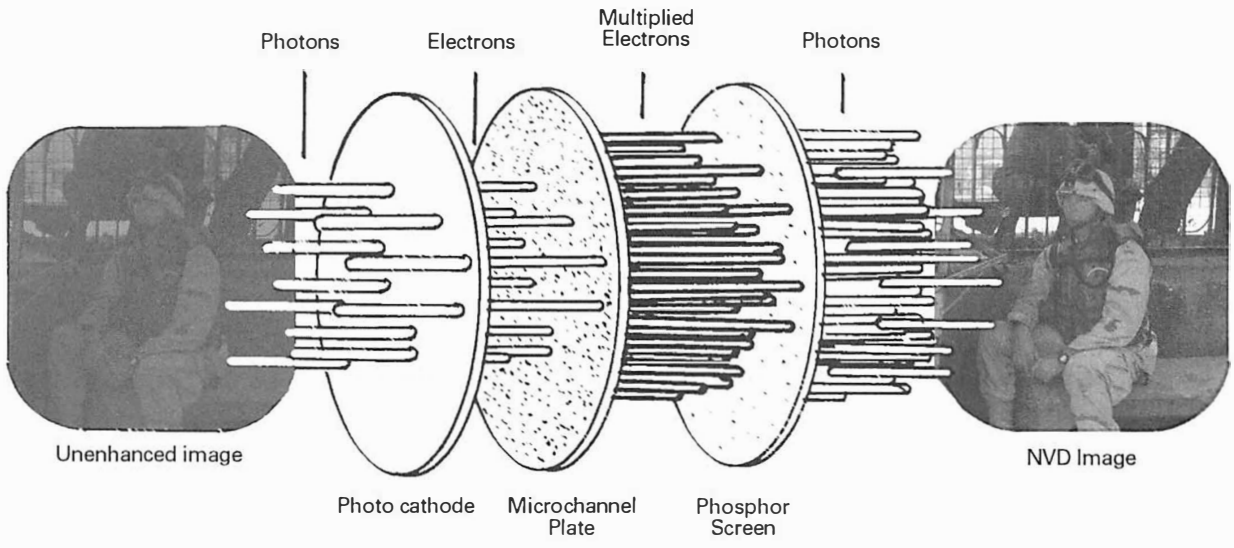


FIGURE 8-45 *The image intensifier tube changes photons to electrons and back again.*

[i] Driggers, 5

[ii] Driggers, 2

[iii] Georgia Tech

[iv] Georgia Tech

[v] Georgia Tech

[vi] Georgia Tech

[vii] Georgia Tech

[viii] Adapted from Hudson, 90

[ix] Pratt

[x] http://www.x20.org/thermal/new.htmPratt

[xi] Georgia Tech

[xii] Hong

[xiii] Georgia Tech

[xiv] Georgia Tech

[xv] Georgia Tech

[xvi] Anderberg

[xvii] Anderberg

[xviii] Driggers, 3

[xix] Hall, 159

[xx] Hall, 162

[xxi] Driggers

[xxii] USN/USMC

[xxiii] USN/USMC, 4–4

[xxiv] USN/USMC

[xxv] Driggers, 194

[xxvi] Driggers, 196

[xxvii] Driggers, 197

[xxviii] Driggers, 198

[xxix] Driggers, 200

[xxx] Driggers, 201

[xxxi] Gurvine, 5–21

[xxxii] Frieden, 294

[xxxiii] Gurvine, 5–21

[xxxiv] Frieden, 303

[xxxv] USN/USMC

[xxxvi] USN/USMC, 1–2

[xxxvii] USN/USMC

[xxxviii] USN/USMC

[xxxix] USN/USMC

[xl] USN/USMC

[xli] USN/USMC

[xlii] USN/USMC, 1–3

9 Principles of Underwater Sound

Introduction

The effectiveness of the modern submarine is based upon two key attributes. Stealth is its primary defensive attribute, since continuous concealment of a submarine prevents the employment of most weapons against it. Its primary offensive attribute, the one that makes it unique among other maritime weapon platforms, is its superior ability to employ acoustic sensors and weapons as a result of its ability to exploit the acoustic environment. A thorough understanding of the acoustic environment is essential in all aspects of submarine warfare.

Before a submarine can be attacked, it must be detected and its location determined within the requirements of the particular weapon system. Detection and localization can take place in two ways. The prosecuting platform may emit energy that is subsequently reflected by the submarine and detected by the sensor. Such sensors may emit light, radar waves, or sound and are known as active sensors.

Conversely, the prosecuting platform may detect energy being radiated from the submarine or detect changes in some aspect of the submarine's environment due to the submarine's presence. Examples include radiation of sound waves, changes in the local magnetic field, or excitation of bioluminescence. Such sensors are known as passive sensors.

The form of energy to be used in an underwater sensor is determined by three factors:

1. Propagation range, or how far that form of energy will travel
2. Resolution or the ability to differentiate between various objects
3. Propagation speed

Light has excellent resolution and high propagation speed; however, its range in water is very limited, on the order of tens of meters. As a result, its operational usefulness in water is limited. This is not to say that light will never be used in undersea warfare; submarines operating at shallow depth are frequently visible from above, and development of optical and laser-based detection systems continues.

Other electromagnetic frequencies, such as radio and radar, have properties similar to light, but sea water is essentially impervious to them at most frequencies. At the low end of the electromagnetic spectrum, where underwater propagation ranges become significant, the lack of resolution caused by the excessively large wavelength, as well as the excessive size of the necessary transmitting equipment and antenna, severely limit the usefulness of this phenomenon as a detection tool.

Magnetic and gravitational field distortions are detectable only at very short ranges because the anomaly diminishes proportionally with the inverse of the range cubed. Although detection range is greater than that of either light or radar, it is only on the order of hundreds of meters and therefore is insufficient for large-area surveillance. However, magnetic anomaly detection (MAD) systems are widely used as localization sensors to pinpoint the location of a submarine once its approximate location has been determined by longer-range/lower-resolution systems.

Other nonacoustic phenomena such as chemical, thermal, or radiation signatures have also been investigated but have not proved to be operationally significant.

Although acoustic energy travels more slowly than electromagnetic radiation, its speed is still considerable. Its propagation range in water is superior to that of all other forms of energy. Although sound generally lacks the resolution of light and other high-frequency electromagnetic radiation, it provides adequate resolution for most underwater sensors.

Because of these attributes, sound continues to be the physical phenomenon of choice in undersea warfare. However, sound is not a panacea. It too has significant

limitations to its effective employment, all of which must be thoroughly understood by the operators of underwater sound equipment.

Fundamental Concepts

Acoustic Waves

All sound, whether produced by a cowbell or a complicated electronic device, behaves in much the same manner. Sound originates as wave motion by a vibrating source and requires an elastic media such as air or water for its transmission. For example, consider a piston suspended in one of these media. As the piston is moved forward and backward, the medium is compressed on the forward stroke and decompressed or rarefied on the return stroke. Thus, a wave motion or series of compressions and rarefactions is caused to move from the source out through the medium.

In the fluid medium, the molecular motion is back and forth, parallel to the direction of the piston's movement. Because the fluid is compressible, this motion results in a series of changes in pressure, defined as "force per unit area." This series of compressions and rarefactions produced by the piston constitutes a compressional wave train. The parameter of an acoustic wave, which measures the disturbance, is therefore pressure.

Another way of understanding the phenomenon of acoustic waves is to consider a string of objects, like weights connected by springs. There are two possible ways of propagating a disturbance down the string; see figure 9-1. The first method is to displace the weight in a direction perpendicular to the string. This will create a transverse wave. As the weight is moved off the axis, the springs from the adjacent weight will exert a restoring force that will pull the weight back on axis.

The other method is to displace the weights along the axis of the string. This creates a longitudinal wave. Again, the restoring force will tend to push the weight back into place. Sound acts like a longitudinal wave and can be modeled as masses attached to springs.[i]

Propagation Speed

The sine wave of figure 9-2 is a graphical representation of the compressional wave train. As the wave passes a given point, the fluid elements are compressed and extended in the manner depicted by the sine wave's oscillations above and below the static pressure. The compressions and rarefactions are labeled on the curve.

There are two important things to note on this curve. The first is the maximum amplitude of the sine wave, labeled *P*, which represents the maximum pressure excursion above and below the static or hydrostatic pressure that exists in the fluid at the location of the wave train.

The second property to note is that the wave train would be propagating at the speed of sound, *c*, in the medium. The units for the speed of sound are meters per second. The relationship between the three acoustic quantities that can be derived from figure 9-2 is:

$$\text{frequency}(f) = \frac{\text{speed}(c)}{\text{wavelength}(\lambda)} \qquad (9\text{-}1)$$

This is the same relationship for electromagnetic energy described in chapter 2.

One final point should be made about the sine wave representation of figure 9-2. Though somewhat difficult to imagine and more difficult to describe pictorially, the displaced parallel lines in figure 9-2 represent the motion of the elements within the field as the wave train passes. As the elements are compressed and extended, their motion can also be mathematically described by a sine wave; however, the elements would be oscillating back and forth about their static position. The maximum amplitude would then be the maximum displacement from the static position.

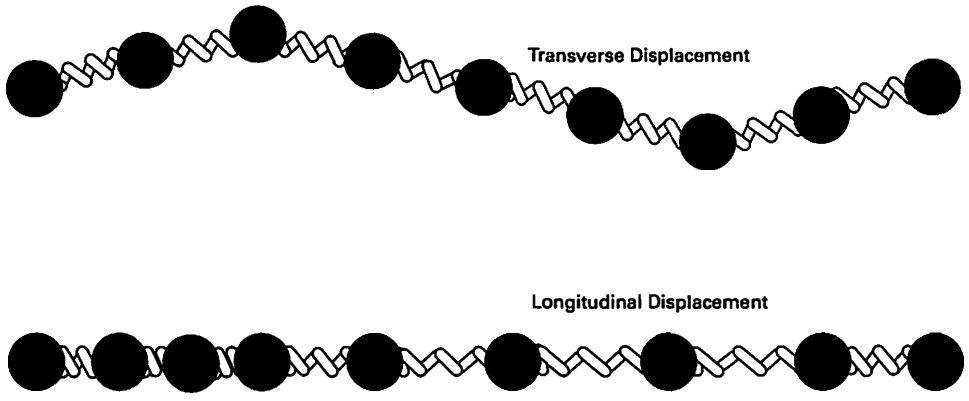


FIGURE 9-1 *Spring and mass model of wave motion*

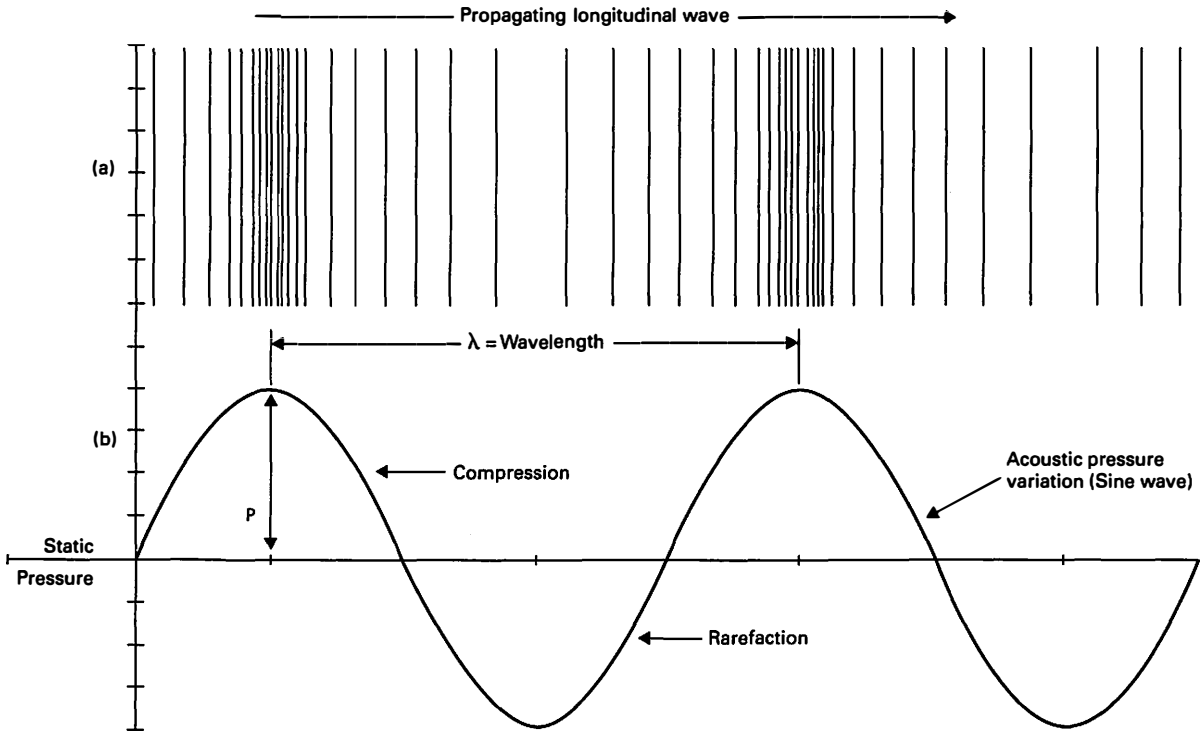


FIGURE 9-2 *Pictorial representation of a simple longitudinal wave*

As an example of the order of magnitude of these displacements, consider that the faintest 1,000 Hz tone that can just barely be heard by humans in air has pressure variations of only 2/10,000,000,000 of one atmosphere of pressure. The corresponding particle displacement is about 10^{-9} centimeter. By way of comparison, the diameter of an atom is about 10^{-8} centimeter.

As the pressure disturbance propagates through the medium, pressure at any point in the medium—as with electromagnetic propagation (see equation 9-2)—can be expressed as a function of distance, r, from the source and the amount of time, t, since a particular wave departed the source.

$$P(r,t) = P(r)\sin\left[\frac{2\pi}{\lambda}(r - ct)\right] \qquad (9\text{-}2)$$

Note how the maximum amplitude, $P(r)$, is dependent on the distance from the source. Neglecting wave interference effects, the amplitude of a pressure disturbance will always diminish; equation 9-2 therefore represents a decreasing amplitude sinusoid wave. If, however, the distance is fixed, then pressure is solely a function of time, and equation 9-2 simplifies to:

$$P(t) = P_A \sin[2\pi(ft)] \qquad (9\text{-}3)$$

where P_A is now the maximum amplitude at the range, r.

Intensity

A propagating sound wave carries mechanical energy in the form of kinetic energy of the particles in motion and potential energy of the stresses set up in the elastic medium. Because the wave is propagating, a certain amount of energy per second—or power—is crossing a unit area, and this power per unit area—or power density—is called the intensity, I, of the wave. As will be shown, the intensity is proportional to the square of the acoustic pressure: $I \propto P^2$. Before giving the defining relationship for intensity, however, two variables must be explained.

The first variable is the root mean square of the pressure. The value of peak pressure, P, as shown in figure 9-2 is not the "effective" or root-mean-square (rms) pressure. An analogy exists between acoustic pressure and the voltages measured in AC circuits. Most voltmeters read the rms voltage. The rms value of a sinusoidal voltage is simply the peak voltage divided by the square root of 2. For example, the common 115-volt line voltage has a peak value of about 162 volts. In order to obtain the rms pressure, P must likewise be divided by the square root of 2.

$$P_e = P_{rms} = \frac{P}{\sqrt{2}} \qquad (9\text{-}4)$$

In this text, the effective or rms pressure as measured by a pressure-sensitive hydrophone will be labeled "P_e."

The explanation of units for pressure will be fully discussed at a later time.

The second variable that must be explained is the proportionality factor that equates intensity to effective pressure squared. It consists of two terms multiplied together—fluid density, ρ, and the propagation speed of the wave, c. The quantity ρc is called the characteristic impedance; it is this property of a sound medium that is analogous to resistance or impedance in electrical circuit theory, where power equals voltage squared divided by resistance.

Additionally, impedance can be illustrated by a simple example: When two dissimilar media such as air and water are adjacent to each other, the boundary between the two is called a discontinuity. When sound energy is traveling through one medium and encounters a discontinuity, part of the energy will be transferred across the boundary and part will be reflected back into the original medium. The greater the difference between the characteristic impedances, the greater will be the percentage of energy reflected.

The difference between the c values for air and water is approximately 1,100 m/sec. Thus, when sound is traveling through water and reaches the surface, only a small amount is transmitted into the air. Most of the energy is reflected by the air/water boundary back into the water.

Obviously, it is important to maintain a consistent set of units when comparing characteristic impedances, and care must be exercised when examining different sources of acoustic information.

It is now possible to formulate an expression for acoustic intensity, the average power per unit area normal to the direction of wave propagation.

$$I = \frac{P_e^2}{\rho c} \qquad (9\text{-}5)$$

The units of acoustic intensity are normally watts/m^2. The importance of equation 9-5 is that it clearly shows the pressure-dependence of the power-transmitting capacity of a train of acoustic waves. If the rms pressure in the water can be measured, then the sound intensity can be determined.

One way to do this is by means of a hydrophone, an electrical-acoustic device much like a microphone that transforms variations in water pressure into a variable electric voltage. Thus, after appropriate calibration, P_e can be read directly from a voltmeter attached to the output of a hydrophone.

Sound Pressure Levels

A convenient system is needed in order to measure and discuss acoustic parameters. Pressure is defined as "force per unit area." Although many people are familiar with the British units of pounds per square inch (psi), it has long been the convention in acoustics to use metric units, namely, newtons per square meter (N/m^2) or dynes per square centimeter (dynes/cm^2).

Of the two metric units, dynes/cm^2 has been the most commonly used. It has an alternate name, *microbar* (μbar), which is equivalent to approximately 1/1,000,000 of a standard atmosphere. For underwater sounds, a reference pressure of 1 μbar was established from which all others are measured. The corresponding reference pressure for airborne sounds is 0.0002 bar, because this is the approximate intensity of a 1,000 Hz tone that is barely audible to human ears.

Of more recent use is N/m^2, also known as a Pascal (Pa). The reference standard derived from this is the micropascal (μPa), which is equivalent to 10^{-6} N/m^2. Table 9-1 provides a summary.

TABLE 9-1 *Acoustic reference conversion factors*

1 μbar	= 1 dyne/cm^2 = 0.1N/m^2
	= 10^5μPa ≈ 10^{-6} atmosphere
1 μPa	= 10^{-6}N/m^2 = 10^{-5}μbar
	= 10^{-5}dyne/cm^2 ≈ 10^{-11} atmosphere

In theoretical investigations of acoustic phenomena, it is often convenient to express sound pressures in newtons/m^2 and sound intensities in watts/m^2. However, in practical engineering work it is customary to describe these same quantities through the use of logarithmic scales known as sound pressure levels.

The reason is related, in part, to the subjective response of the ear. The human ear can hear sounds having pressure disturbances as great as 100,000,000 micropascals and as small as 10 micropascals. A problem is encountered when discussing pressures that vary over so great a range, in that the minimum audible disturbance is one ten-millionth that of the maximum.

In underwater acoustics, useful pressures having even greater variations in magnitude are commonly encountered. In order to make the numbers more manageable, both in magnitude and for actual manipulation, logarithms are used rather than the numbers themselves.

EXAMPLE 9-1: Two acoustic signals are to be compared, one having a P_e of 100,000,000μPa and the other a P_e of 10μPa. Their ratio would be:

$$\frac{P_1}{P_2} = \frac{100{,}000{,}000\mu\text{Pa}}{10\mu\text{Pa}} = 10{,}000{,}000 = 10^7$$

In underwater acoustics, however, the attribute of primary interest is sound intensity, or power, rather than pressure. As with pressure, acoustic intensities are referenced to some standard intensity, designated I_0, and the logarithm of the ratio taken. Intensity level (IL) is therefore defined as:

$$IL = 10 \log\left(\frac{I}{I_0}\right) \qquad (9\text{-}6)$$

where IL is measured in dB.

However, because there is only one intensity reference (10^{-12} watt/m^2 in air) and many pressure references, IL must be able to be expressed in terms of pressure. By inserting equation 9-5 into equation 9-6, a new expression of IL can be obtained that is based on pressure rather than on intensity per se.

$$IL = 10 \log\left(\frac{\frac{P_e^2}{\rho c}}{\frac{P_0^2}{\rho c}}\right)$$

or (9-7)

$$IL = 10 \log\left(\frac{P_e^2}{P_0^2}\right) = 20 \log\left(\frac{P_e}{P_0}\right)$$

Under the assumption that the reference intensity and the reference pressure are measured in the same acoustic wave, then a new sound level called sound pressure level can be defined.

$$SPL = 20 \log\left(\frac{P_e}{P_0}\right) \qquad (9\text{-}8)$$

SPL has the dimensionless units of decibels. Since the voltage outputs of the microphones and hydrophones commonly used in acoustic measurements are proportional to pressure, acoustic pressure is the most readily measured variable in a sound field. For this reason, sound pressure level is more widely used in specifying sound levels, and this is also why only pressure references are used in underwater acoustics. Note that IL and SPL are numerically equivalent.

$$IL = 10 \log\left(\frac{I_e}{I_0}\right) = 20 \log\left(\frac{P_e}{P_0}\right) = SPL \qquad (9\text{-}9)$$

where:

$$I_0 = \frac{P_0^2}{\rho c}$$

Decibels

The decibel system was selected by acousticians for a number of logical reasons. First, it is a logarithmic system, which is convenient for dealing with large changes in quantities. Logarithms also simplify computations since multiplication and division are reduced to addition or subtraction, respectively.

Second, human senses have an approximate logarithmic response to stimuli such as light, sound, and heat. For example, the human ear perceives about the same change in loudness between one and ten units of pressure as it perceives between ten and hundred units of pressure.

And finally, in the area of underwater acoustics, the primary interest is in ratios of power levels and signal levels rather than absolute numerical values.

In the decibel system, the bel is the fundamental division of a logarithmic scale for expressing the ratio of two amounts of power. The number of bels used to express such a ratio is the logarithm to the base 10 of the ratio. Acousticians decided the bel was too large a unit for application in their field, and subsequently adopted the decibel (1/10 bel) as their basic logarithmic unit.

The conversion factors in table 9-1 can in themselves be cumbersome to use, but when expressed in dB, only addition or subtraction is required. When converting from a pressure referenced to 1 bar to a pressure referenced to 1 Pa, simply add 100dB. When converting from 0.0002 bar to 1 Pa, simply add 26dB. When converting from 1 Pa to the others, merely subtract the appropriate values.[ii]

As an aid in interpreting and understanding the decibel scale and its relation to intensity and pressure, it is useful to remember that:

a factor of 2 in intensity is +3dB
a factor of 0.5 in intensity is −3dB
a factor of 10 in intensity is +10dB
a factor of 0.1 in intensity is −10dB
a factor of 2 in pressure is +6dB
a factor of 0.5 in pressure is −6dB
a factor of 10 in pressure is +20dB
a factor of 0.1 in pressure is −20dB

In understanding intensity levels and sound pressure levels, it is important to note that the decibel scale is a ratio of power or energy, no matter what quantities are being compared. A problem commonly arising in acoustic calculations is obtaining the overall intensity level after the individual intensities within the applicable bandwidths have been calculated.

Such a situation is encountered in calculating a term in the sonar equations (to be discussed later) called noise level,

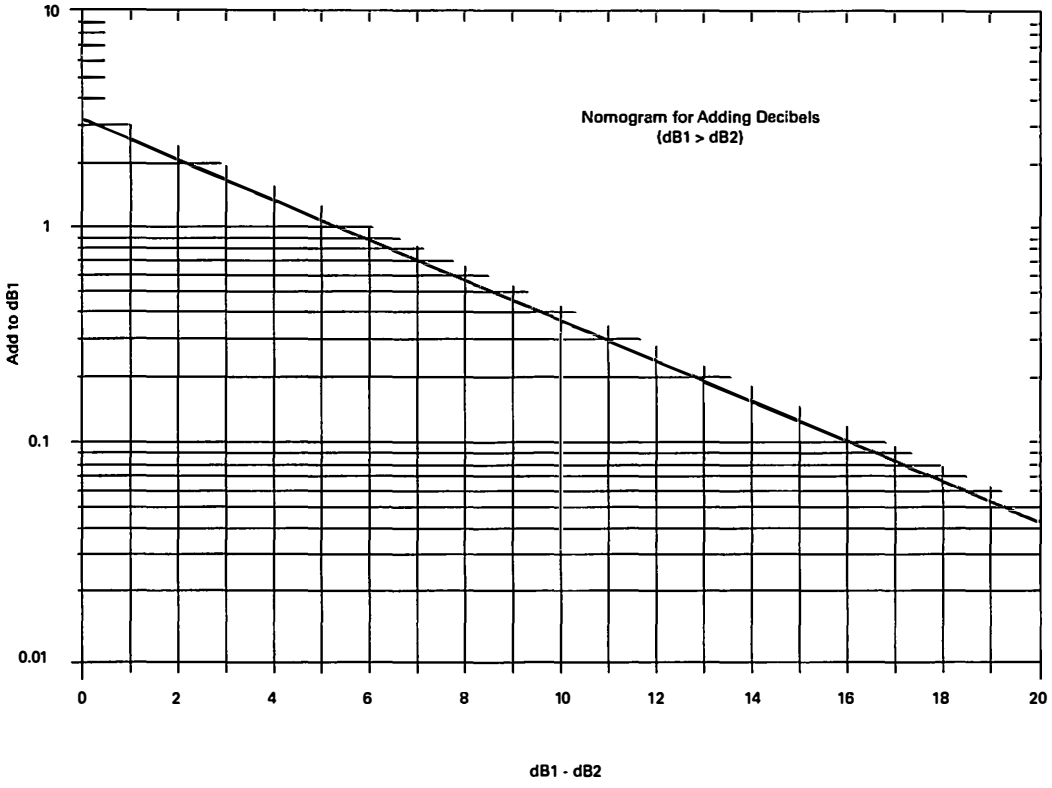


FIGURE 9-3 *Nomogram to combine decibels*

which is actually a combination of ambient noise and self-noise. When working with decibels, it is not possible to merely add similar intensity levels together and work with their sum.

For example, two 30dB signals combine to give a total intensity level of 33dB, not 60dB as might be expected. The reason for this is that, as shown above, doubling the intensity is represented in decibels by a +3dB change.

The process is more complicated when dealing with levels of unequal value. Figure 9-3 can be used to determine the dB increase above a level of dB1, in terms of the difference dB1 − dB2, to be expected when dB1 and dB2 are combined. This process can be expanded to include any number of intensity levels.

However, when combining more than two intensities, it is often easier to use antilogs. Antilogs convert each intensity level back to its intensity units, which are then added together and then reconverted back to dB levels. Either method may be used and should result in the same numerical value of noise level.

EXAMPLE 9-2: Three SPLs at pressures of 63, 64, and 66dB are received. What is the total SPL in dB when combined?

SOLUTION: *NOMOGRAM METHOD*

Step 1:

$$66\text{dB} - 64\text{dB} = 2\text{dB}$$

Entering the nomogram at 2dB gives a corresponding value of 2dB, which is then added to 66dB.

$$66\text{dB} + 2\text{dB} = 68\text{dB}$$

Step 2:

$$68\text{dB} - 63\text{dB} = 5\text{dB}$$

Entering the nomogram at 5dB gives a corresponding value of 1.2dB, which is then added to 68dB.

$$68\text{dB} + 1.2\text{dB} = 69.2\text{dB}$$

Therefore:

$$63\text{dB} + 64\text{dB} + 66\text{dB} = 69.2\text{dB}$$

It can be shown that order does matter with this method; therefore a range of values will result.

ANTILOG METHOD

As described, this is a much more accurate method but requires a calculator or slide rule.

$$\text{Total SPL} = 10\log\left(10^{63/10} + 10^{64/10} + 10^{66/10}\right)$$
$$= 69.28\text{ dB}$$

Although standardization has been reached for measuring intensities, such is not the case for other quantities. Ranges are expressed in yards, kilometers, and nautical miles. Depths are given in feet, meters, or fathoms. Sound speed is stated in feet per second or meters per second

and ship speed in knots. Temperatures are commonly specified in degrees Fahrenheit or degrees Celsius.

Because of these diverse units, the user should exercise due caution when discussing the various facets of underwater sound to ensure that misunderstandings do not occur. In this text, the SI units are used wherever possible.

Speed of Sound in the Sea

From physics, we know that when gas is the transmitting medium, the denser the gas, the slower the speed of sound. Yet the speed of sound in water is about four times greater than that in air. The reason for this apparent contradiction is that there is another, more important factor that influences the speed of sound. As it turns out, the speed of sound is determined primarily by the elasticity of the medium, and to a lesser extent by its density.

Elasticity is defined as "the property of a body that causes it to resist deformation and to recover its original shape and size when the deforming forces are removed." Of specific concern is volume elasticity or bulk modulus (β), that is, the ratio of force per unit area (stress) to the change in volume per unit volume (strain).

$$\beta = \frac{\text{stress}}{\text{strain}} \qquad (9\text{-}10)$$

In order to change the volume of a liquid, it is necessary to exert a force of much greater magnitude than is required to bring about an equivalent change in the same volume of air. Therefore, the value of bulk modulus is much greater for a liquid than for a gas. The speed of sound, c, is equal to the square root of the ratio of bulk modulus to density:

$$c = \sqrt{\frac{\beta}{\rho}} \qquad (9\text{-}11)$$

where:

c = speed of sound in medium
β = bulk modulus
ρ = density

Although sea water is almost a thousand times denser than air, the enormous bulk modulus of water is the more important factor determining sound speed. Of concern, however, are not the differences of the two media but the conditions in water that cause changes in sound speed.

Contrary to the assumptions made with electromagnetic energy and the speed of light in the atmosphere, the ocean is not a homogeneous medium, and the speed of sound varies from point to point in the ocean. This variation in sound speed is one of the most important characteristics affecting the transmission of sound. The three main environmental factors affecting the speed of sound in the ocean are salinity, pressure, and temperature.

Salinity, which on average ranges from 32 to 38 parts per thousand (ppt), is fairly constant in the open ocean. A change of salinity will cause a small corresponding change in density with a resulting change in bulk modulus, causing variation of sound speed. The greatest variation in salinity in the open ocean exists in the vicinity of "oceanic fronts," which are narrow zones separating water masses of different physical characteristics and usually exhibiting very large horizontal gradients of temperature and salinity.

Even greater variation in salinity can be expected around the mouths of rivers, in heavy ice, and in areas of extraordinary rainfall—the monsoon, for example—where a layer of fresh water overrides a layer of salt water. A change in salinity of 1 ppt will result in a change in sound speed of approximately 1.3 meters per second.

Pressure in most circumstances is more important than salinity, but in the sea its change is constant and thus predictable. It also causes a change in bulk modulus and density, and the result is an increase in sound speed of 0.017 m/sec for every meter of depth increase. This slight change, which is important when temperature remains constant, causes sound to bend upward at great depths, as will be discussed later.

Temperature, the predominant factor affecting sound speed, usually decreases with depth. This leads to an accompanying decrease in sound speed at the rate of approximately 3 meter/sec per degree Celsius. Below a depth of about 1,000 meters, however, temperature is fairly constant, and the predominant factor affecting sound speed becomes pressure.

At first glance it would seem that a temperature decrease would increase sound speed, due to the increased water density, but this is not so. As the temperature of a medium decreases, bulk modulus decreases while density increases. Considering these effects in terms of the sound speed formula in equation 9-11, it is clear that a decrease in temperature brings an attendant decrease in sound speed.

It also should be noted that the effect of temperature changes on bulk modulus and density differs at a variable rate. A change in temperature at one point on the scale, therefore, affects sound speed differently than an equal change at another point on the scale.

Additionally, the effect of temperature is relatively large compared to the effects of other factors. It takes a depth change of about 165 meters to cause the same change in sound speed as a 1° temperature change. As will be discussed, temperature is therefore the only factor normally measured and evaluated under operational conditions.

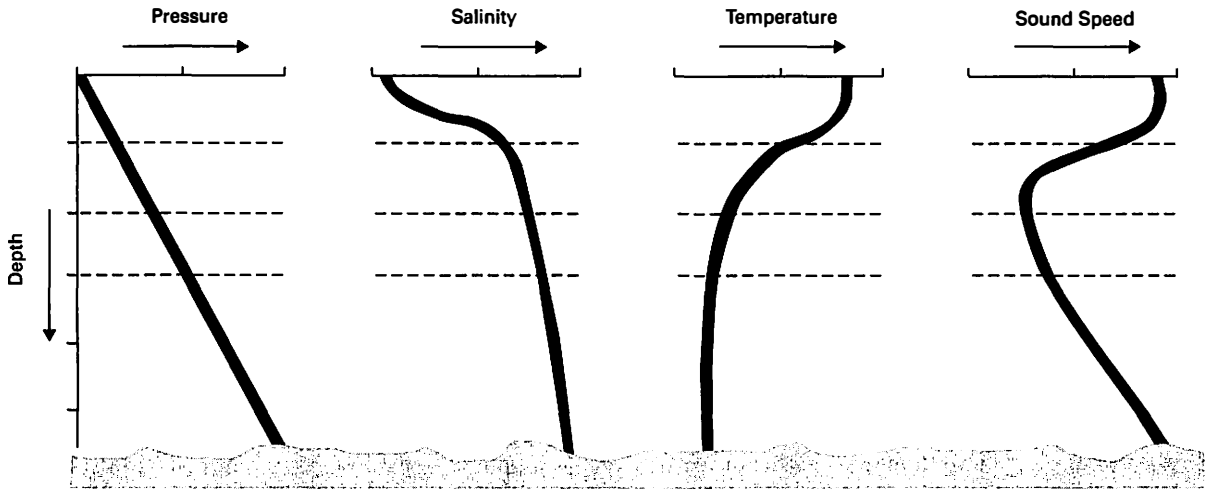


FIGURE 9-4 *Graphical relationship of sound speed to pressure, salinity, and temperature*

The relationship relating frequency, wavelength, and velocity still holds, namely:

$$f\lambda = c_{(T,p,S)} \tag{9-12}$$

where:

$c_{(T,p,S)}$ = speed of propagation as a function of temperature (T), pressure (p), and salinity (S)

Dealing with these three factors to arrive at values for bulk modulus and density, and thence sound speed, is very cumbersome. To overcome this, numerous empirical relationships have been developed for converting the three factors directly to sound speed. The speed of propagation has a fairly complicated dependence on these variables, and has been determined empirically.

A simplified version of such sound speed equations, developed by Mackenzie in 1981, is:

$$\begin{aligned} c_{(T,p,S)} = {} & 1{,}448.96 + 4.59T - 5.304\times10^{-2}T^2 \\ & + 2.374\times10^{-4}T^3 + 1.340(S-35) \\ & + 1.630\times10^{-2}D + 1.675\times10^{-7}D^2 \\ & - 1.025\times10^{-2}T(S-35) - 7.139\times10^{-13}TD^3 \end{aligned} \tag{9-13}$$

where:

c = m/s
T = temperature in degrees Celsius
S = salinity in parts per thousand
D = depth in meters[iii]

By way of contrast, the equation for the speed of sound in air is approximately

$$c_{(T)} = 331.6 + 0.6T$$

In making calculations involving the transmission of sound through the sea, it is sometimes adequate to use a standard speed rather than the more accurate value given in equation 9-13. Although in sea water c can vary from a low of about 1,420 m/s to over 1,560 m/s depending on environmental conditions, a standard speed of 1,500 m/s may be assumed for computation purposes unless otherwise noted.

Field Observations of Sound Speed

Knowledge of sound velocity is important to the undersea warfare tactician and physical oceanographer because of the effect that variations in sound velocity have upon acoustic refraction. Devices in use today to estimate sound velocity fall into two main catagories.

The first device is called a bathythermograph (BT). As previously stated, temperature is the predominant variable affecting sound speed in the ocean. Not only is it relatively easy to measure, but when applied to empirical relationships such as equation 9-13, sound speed can be computed.

Older BT systems employed a mechanical device that was lowered on a cable and the temperature was scribed on a smoked piece of glass. This had a number of inherent disadvantages that have been overcome through the development of the expendable bathythermograph (XBT), which does not require retrieval of the sensing unit; see figure 9-5.

A diagram view of an XBT is shown in figure 9-6. It consists of a thermistor probe that is ejected from the launching platform and sinks at a known rate. The XBT is connected by a fine wire to a special recorder on board the launching platform. As it sinks, the thermistor's electrical resistance changes with changing temperature, and a temperature versus depth trace is obtained.

Because the wire uncoils from both the probe and its launcher, there is no tension on the wire until the probe

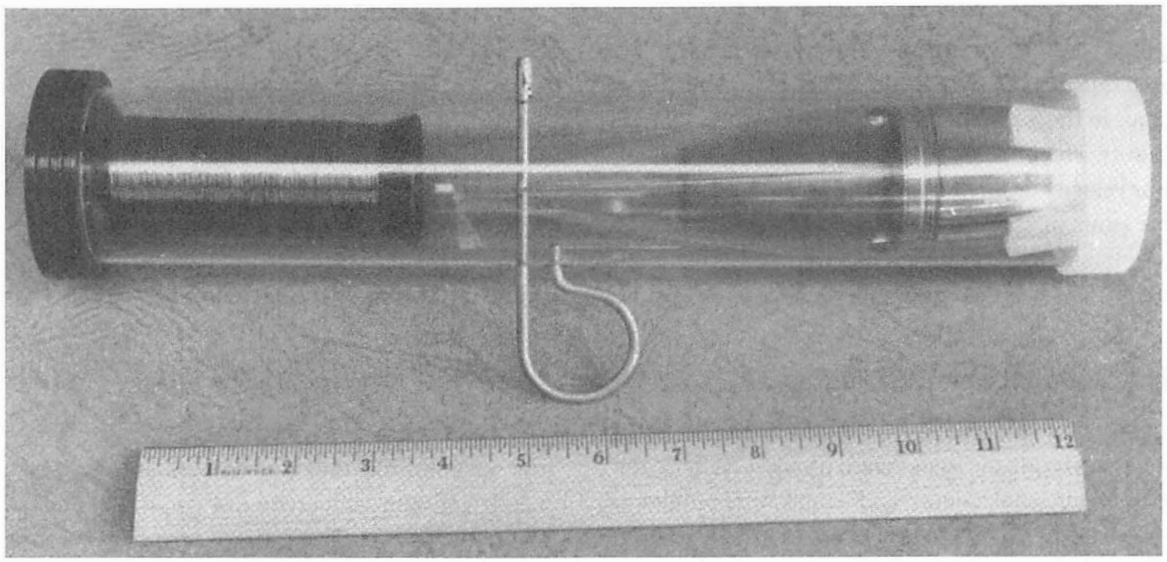


FIGURE 9-5 *Expendable bathythermograph probe*

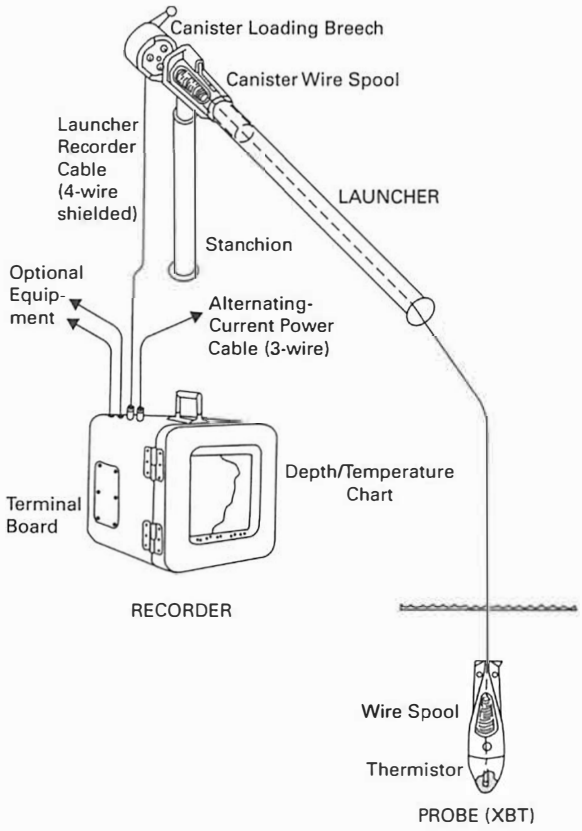


FIGURE 9-6 *Expendable bathythermograph*

has sunk to its full depth. At that point, the wire snaps and the recording stops. Variants of the basic XBT have been developed for use from aircraft through incorporation into sonobuoys and aboard submarines.

When the XBT temperature versus depth trace is converted to sound speed versus depth, it produces a sound velocity profile that is of sufficient accuracy for most operational requirements.

The second and most accurate method is the sound velocimeter. Its principal advantage is that it can measure sound speed directly, without the need for conversions, by transmitting a pulse of sound over a very short path on the order of 1/2 meter or less. When the pulse arrives at the receiver, another pulse is then triggered from the transmitter.

This is known as the "sing-around" or "howler" principle. The faster the speed of sound in the water in which the velocimeter is submerged, the faster the pulse will travel and the sooner it will arrive at the receiver to trigger the succeeding pulse. Since nearly all the time delay between pulses occurs as acoustic delay in the water, the pulse repetition frequency (PRF) is directly proportional to the local sound speed. Thus, knowing the path length and observing the PRF can lead directly to computation of sound speed.

Until recently, sound velocimeters were expensive and awkward to use, thus eliminating their use tactically. The recent development of the expendable sound velocimeter (XSV) has made it possible to reduce sound velocity measurement errors to less than 0.25 meter per second at reasonable expense without reduction of the mobility of combatant units. Today's sophisticated sonars and acoustic navigation systems can provide improved information in many oceanic regions when actual sound velocity profiles are used rather than extrapolated sound velocity values based on temperature profiles and assumed salinity data.

Based on the variability in the sea with time, a policy of regular, periodic measurement of the velocity profile is required during underwater operations. Normally, one or two ships in the force are assigned the bathythermograph guard duty. These ships periodically measure the temperature or velocity profile, and promulgate it to all undersea warfare units in company. These sound velocity profiles are essential in determining the sound propagation paths available.

Typical Sound Speed Profiles

It is important to remember that although temperature is the dominant factor, the sound speed profile is really a composite of the pressure, salinity, and temperature profiles, as shown in figure 9-4. In the area of ocean fronts, where salinity may vary up to 3 ppt from assumed values, the use of temperature data alone may result in an error of up to 4.2 meters per second in the calculation of sound speed.

A typical composite deep-sea sound speed profile is shown in greater detail in figure 9-7. The profile may be divided into four major layers, each having different thermal characteristics. Just below the sea surface is the surface layer, in which the speed of sound is susceptible to daily and local changes of heating, cooling, and wind action. The surface layer may contain isothermal water that is formed due to mixing by wind as it blows across the water.

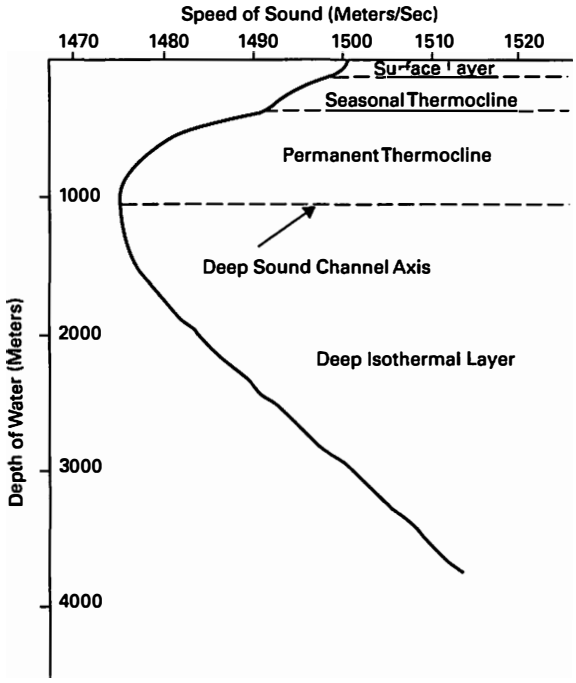


FIGURE 9-7 *Typical deep-sea speed profile divided into layers*

Below the surface layer lies the seasonal thermocline. The word *thermocline* denotes a layer in which the temperature changes rapidly with depth. The seasonal thermocline is characterized by a negative sound speed gradient that varies with the seasons. During the summer and fall, when the near-surface waters of the sea are warm, the seasonal thermocline is strong and well defined; during the winter and spring, and in the Arctic, it tends to merge with and be indistinguishable from the surface layer.

Underlying the seasonal thermocline is the permanent thermocline, which is affected only slightly by seasonal changes.

Below the permanent thermocline and extending to the sea bottom is the deep isothermal layer, which has a nearly constant temperature of about 4°C and in which the speed of sound has a positive gradient because of the effect of pressure on sound speed.

Between the negative speed gradient of the permanent thermocline and the positive gradient of the deep isothermal layer, there is a speed minimum toward which sound traveling at great depths tends to be bent or focused by refraction. This is the deep sound channel, which will be discussed later.

The refraction of sound, however, is much more complex than this simple four-layer ocean model would indicate. There are ocean eddies, fronts, interfaces between currents, underwater mountains, and ridges. For instance, a submarine detected in the Labrador Current but crossing into the Gulf Stream has been compared to a person going out into an open field and then disappearing into the nearby woods.[iv]

Ray Theory

The propagation of sound in an elastic medium can be described mathematically by finding solutions of the wave equation using the appropriate boundary and medium conditions for a particular problem. The wave equation is a partial differential equation relating the acoustic pressure P to the coordinates x, y, and z and the time t, and may be written as:

$$\frac{\partial^2 P}{\partial t^2} = c^2 \left(\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2} + \frac{\partial^2 P}{\partial z^2} \right) \qquad (9\text{-}14)$$

There are two theoretical approaches to a solution of the wave equation. In one, called normal-mode theory, the propagation is described in terms of characteristic functions called normal modes, each of which is a solution of the equation. The normal modes are combined additively to satisfy the boundary and source conditions of interest.

The result is a complicated mathematical function that, though adequate for computations on a digital computer, gives little insight on the distribution of the energy of the source in space and time. Compared to ray theory, normal-mode theory is well suited for a description of sound propagation in shallow water, but will not be used in this text.

The other solution of the wave equation is ray theory. The essence of ray theory is (1) the postulate of wave fronts, along which the phase or time function of the solution is constant, and (2) the existence of rays that describe where in space the sound emanating from the source is being sent. Like its analog in optics, ray acoustics has considerable intuitive appeal and presents a picture of the propagation in the form of the ray diagram.

For almost all operational problems, with the exception of near fronts and eddys, the sound speed gradient with respect to horizontal changes of location can be assumed to be zero. The major gradient of interest is the vertical gradient, *dc/dz*, where *dz* is the amount of depth change. If a source of sound at the surface of the sea radiates omnidirectionally, a wave front expanding from this source in all directions transfers energy from one particle to another, and by this means, the wave is propagated. If some point on this wave front is selected and a line drawn from it in the direction of energy propagation, then connecting these points as the wave expands in space will result in a line called a ray, as illustrated in figure 9-8.

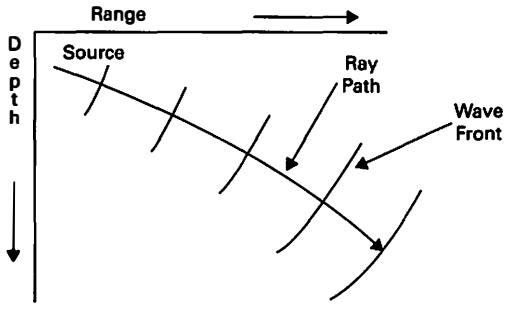


FIGURE 9-8 *Ray theory*

A sound wave, or ray, that enters another medium or layer of the same medium having a different characteristic impedance will undergo an abrupt change in direction and speed. Depending upon the angle of incidence and the abruptness of change in ρc, a portion of the impinging acoustic energy will be reflected off the medium boundary, and a portion will be refracted or bent as it passes through the boundary. A good "rule of thumb" is that a sound ray will always bend toward the region of slower sound speed.

One of the most important practical results of ray theory is Snell's Law, which describes the refraction of

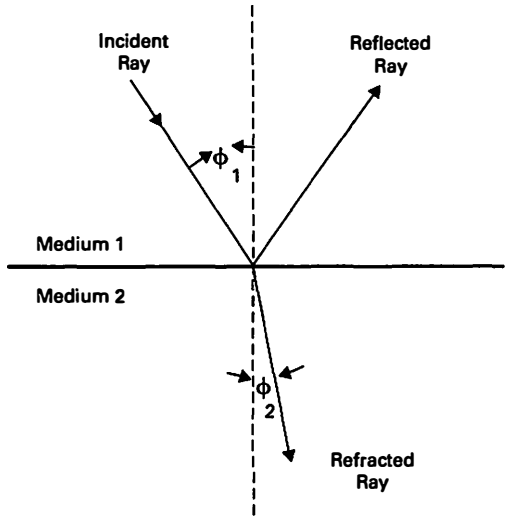


FIGURE 9-9 *Snell's Law*

waves in media of variable speeds; see figure 9-9. Snell's Law states that the angle of incidence, φ_1, at a boundary is related to the angle of refraction, φ_2, by the following expression:

$$\frac{\sin\varphi_1}{\sin\varphi_2} = \frac{c_1}{c_2} \qquad (9\text{-}15)$$

where:

c_1 = sound speed in medium 1
c_2 = sound speed in medium 2

If the wave is considered to be passing through three horizontal layers or strata, in each of which the sound speed is considered to be constant, then Snell's Law can be rewritten as

$$\frac{c_1}{\cos\theta_1} = \frac{c_2}{\cos\theta_2} = \frac{c_3}{\cos\theta_3} = \frac{c_n}{\cos\theta_n} \qquad (9\text{-}16)$$

where:

c_n = speed of sound at any point in the medium
θ_n = angle made with horizontal at that point

Note that the angle in equation 9-16 is the complement of the angle usually expressed in Snell's basic law. It is commonly referred to as the grazing angle or angle of inclination. This expression is the basis of the ray computation used by most computers, since it enables a particular ray to be "traced out" by following it through the successive layers into which the speed profile may have been divided. In a layered medium having layers of constant speed, the rays consist of a series of straight-line segments joined together, in effect, by Snell's Law; see figure 9-10.

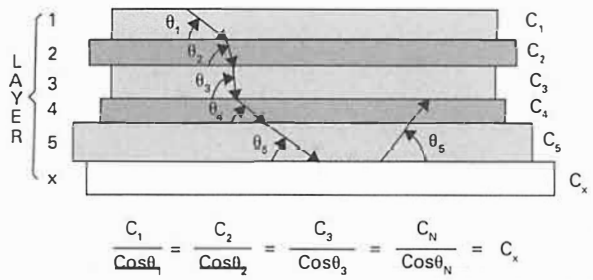


FIGURE 9-10 *Snell's Law applied to multiple layers*

In practice, however, temperature does not change abruptly; rather, the gradient will normally decrease or increase at a measurable rate. For such a situation, the sound speed at any depth z would be given by:

$$c(z) = c_0 + gz \tag{9-17}$$

where:

c_0 = speed at the surface or transducer depth
g = speed gradient dc/dz between the surface and depth z

The net result is that, in reality, ray traces appear as curves rather than straight lines. By combining equations 9-16 and 9-17, an expression can be developed for the radius of curvature R of any ray at any point along the ray path, as shown by equation 9-18 and figure 9-11.

$$R = \left|\frac{c_0}{g}\right| = \left|\frac{c}{g\cos\theta}\right| \tag{9-18}$$

Under operational conditions, values of R are very large, approaching several tens of kilometers.

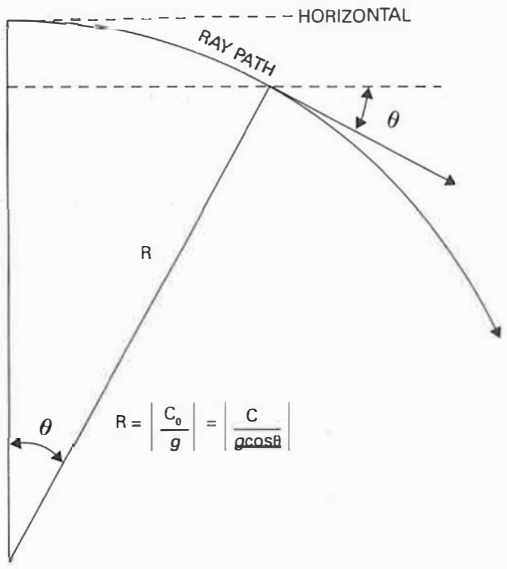


FIGURE 9-11 *Radius of curvature*

Propagation Paths

Thermal Structure

The thermal structure of the ocean governs the refractive conditions for a given water mass. Despite infinite vertical temperature variations in the ocean, the temperature structure normally can be related to three basic types: (1) isovelocity, (2) negative temperature gradient, and (3) positive temperature gradient. In an isothermal condition, the water's temperature is almost constant. If there is a slight decrease in temperature that is just balanced out by the pressure increase, the result is an isovelocity condition. This causes a straight-line ray leaving the source in lines that continue with little or no change in angle. Long-range sound propagations are possible when this type of structure is present; see figure 9-12.

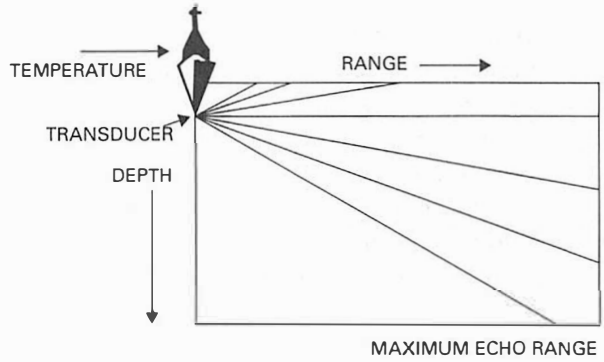


FIGURE 9-12 *Isovelocity condition*

When there is a negative temperature gradient, sound speed decreases with depth, and sound rays bend sharply downward. This condition is common near the surface of the sea. At some horizontal distance from the sound source, beyond where the rays bend downward, is a region in which sound intensity is negligible; see figure 9-13. This region is called a shadow zone.

The magnitude of the temperature gradient determines the amount of bending of the sound beam and thus the range of the shadow zone. For example, if the decrease in temperature to a depth of 10 meters totals 2°C or more, the shadow zone would begin beyond a horizontal range of 1,000 meters due to the sharp curvature of the sound beam.

When the temperature of the water has a positive gradient, sound speed increases with depth, and sound rays are refracted upward; see figure 9-14. Longer sound propagation ranges are attained with this temperature structure than with a negative gradient because the rays are refracted upward and then reflect off the sea surface. Unless the surface of the sea is very rough, most of the

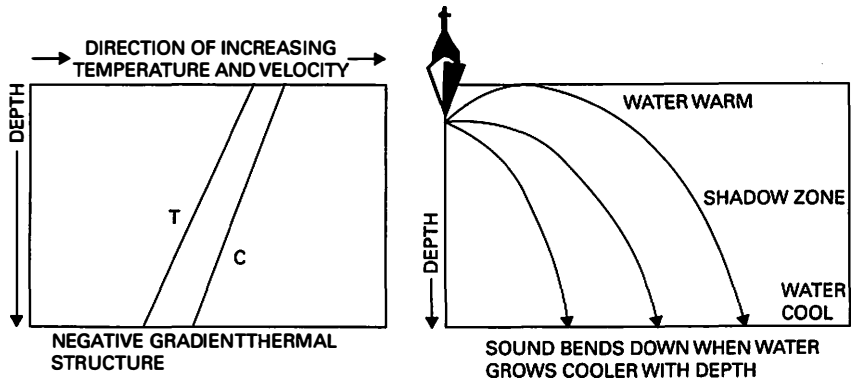


FIGURE 9-13 *Negative velocity gradient*

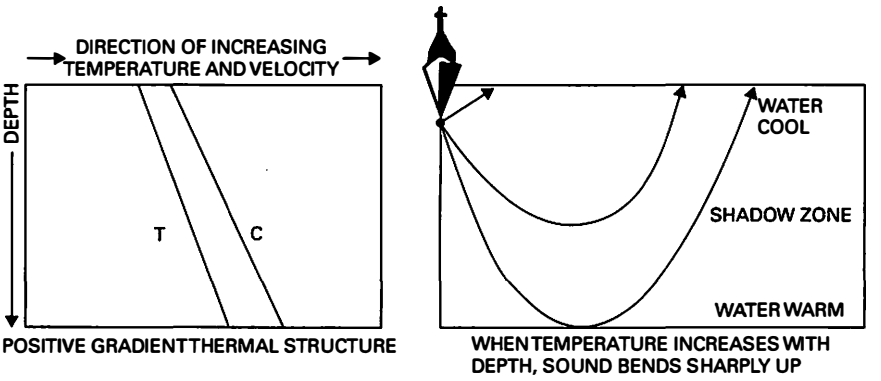


FIGURE 9-14 *Positive velocity gradient*

rays are repeatedly reflected at the highly reflective air/water boundary, resulting in extended ranges.

In discussing sound propagation, it is customary to use the temperature profile as an indicator of sound speed conditions at various depths, because it has the greatest effect. It must be remembered, however, that changes in sound direction result from changes in the sound speed profile, which is influenced not only by temperature but also by pressure and salinity.

Circumstances usually produce conditions where combinations of temperature gradients occur with depth. One of these combinations includes a layer of isothermal water over water with a negative gradient. Approximately 90% of all the bathythermograph records from all over the world show this type of thermal structure. One ray, labeled "the critical ray," becomes horizontal at the boundary or division between the isothermal layer and the negative gradient. The speed of sound is a maximum at this boundary point.

This positive gradient over a negative gradient produces a special kind of propagation where rays split at the boundary, which is called the sonic layer. The depth of maximum sound velocity, which occurs at this layer, is called the sonic layer depth (SLD). Consequently, we define the sonic layer depth (z) as that depth of greatest sound speed (c) in the seasonal thermocline.

One-half of the critical ray bends toward the upper region at a reduced speed, and the other half bends toward the lower region at a reduced speed. The angle that the critical ray makes with the horizontal at the point of projection is called the critical angle. All rays of sound directed at an angle less than the critical angle will follow paths entirely within the isothermal layer and will be bent upward to the surface.

All sound rays directed at an angle greater than the critical angle follow paths that penetrate the boundary and are subsequently refracted downward. No rays enter the region bounded by the two branches of the split critical ray, and for this reason, it is also called a shadow zone. Sharp shadow zones are not fully developed because of diffraction and other effects, though the sound intensity in this area is quite low.

Submarine commanders deliberately use this phenomenon, when it exists, to attempt to escape detection when approaching a target. The optimum depth for close approach to a target with minimum probability of counter-detection is approximately:

$$\text{best depth} = 17\sqrt{z} \qquad (9\text{-}19)$$

where:

z = layer depth in meters

This is accurate down to a layer depth of 60 meters. Below that, the best depth for approach is a constant 60 meters below layer depth.

Direct-Path Propagation

Direct path is the simplest propagation path. It occurs where there is approximately a straight-line path between the source and the receiver, with no reflections and only one change of direction due to refraction; see figure 9-15.

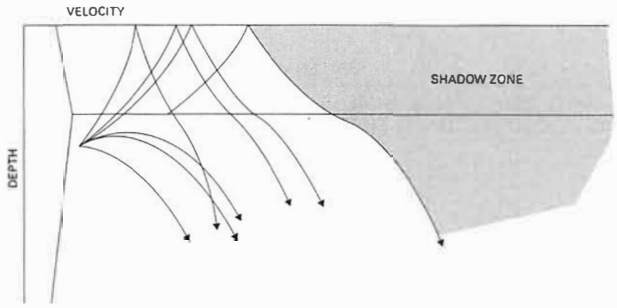


FIGURE 9-15 *Direct-path propagation*

Surface Duct Propagation

In this kind of propagation, the sound source is in a region with a positive velocity gradient. This causes most of the ray paths to refract upward. Unless the surface of the sea is very rough, most of these rays are repeatedly reflected at the surface to produce longer ranges. The overall net effect is to trap the sound in a relatively small layer below the surface; this effect is called a surface duct; see figure 9-16.

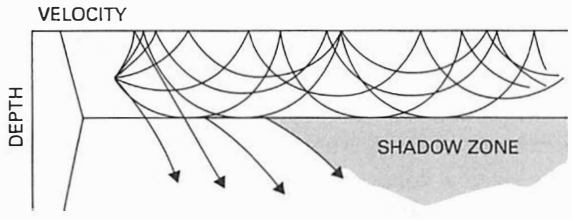


FIGURE 9-16 *Surface duct propagation*

Half-Channel Propagation

When a positive velocity gradient exists from the surface to the bottom, a condition known as half channel occurs. The positive gradient causes all the ray paths to refract upward. This is essentially a surface duct that exists from the surface to the bottom. There are no shadow zones in half-channel propagation, and the maximum direct path ranges are achieved; see figure 9-17.

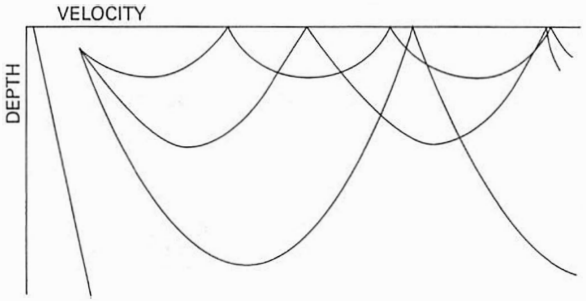


FIGURE 9-17 *Half-channel propagation*

However, like surface ducting, half-channel losses are sea state dependent. Although there is technically no best depth, there is some advantage to the submarine being as shallow as possible. This situation causes the ray paths to reflect at a closer range, adding surface reflection losses to an otherwise purely refracted path.

Sound Channel Propagation

In the deep ocean, temperature usually decreases with depth to approximately 1,000 meters. Deeper than this, temperature is a constant 4°C and sound speed increases as a result of pressure. Negative speed gradient overlays a positive speed gradient, allowing equal speeds at two different depths with slower speed conditions in between. When water with a negative speed gradient overlays a positive speed gradient, a sound channel is produced; see figure 9.18.

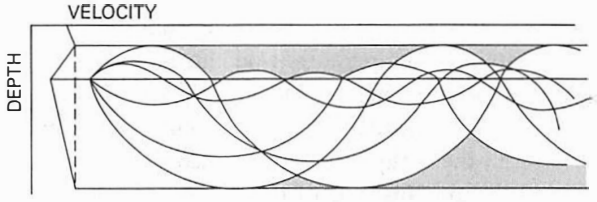


FIGURE 9-18 *Sound channel propagation*

Under these circumstances, any sound signal traveling in this area is refracted back and forth so that it becomes horizontally channeled. Sound rays originating with an initial upward inclination are refracted downward.

Rays from a sound source in this layer that make a small angle with the horizontal are roughly sinusoidal, crossing and recrossing the layer of minimum speed.

This reinforcement of rays within the sound channel can continue until the sound is absorbed, scattered, or intercepted by some obstacle. Sounds traveling in this manner sometimes are received at extremely great distances from the source. These long ranges occur primarily as a result of two factors: (1) absorption is small for low-frequency sound and (2) most of the sound energy from a sound source at the axis is confined to the channel.

The region of minimum sound velocity, at a depth exceeding a kilometer, is the deep sound channel. It acts as an acoustic wave guide; sound propagating either upward or downward encounters sound velocity gradients and bends back into this channel.

Within the deep sound channel, sound undergoes only cylindrical spreading losses, in which intensity drops off only as the first power of the distance. If sound spreads uniformly in all directions, known as spherical spreading, this falloff would follow the square of the distance. The concept of cylindrical and spherical spreading and their importance to sonar is presented in the next chapter. Figure 9-19 shows the ray paths existing for a deep sound channel. Rays A and B are bounded by the sound channel. Other rays may follow propagation paths that are reflected at the sea surface and bottom, or refracted and reflected from either the sea bottom or the sea surface.[v]

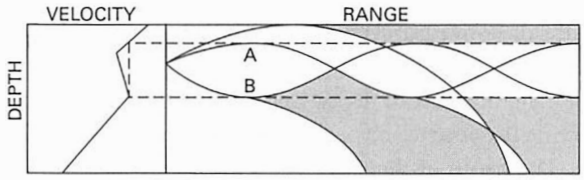


FIGURE 9-19 *Deep sound channel*

Under certain circumstances, a sound channel can exist near the surface of the sea. In a surface layer with a strong positive temperature gradient, the upward bending of sound rays combined with reflections from the surface will form such a channel. Sonar ranges many times greater than normal have been observed where sound channels exist. However, the conditions that produce such sound channels near the surface are rare and not very stable.

Convergence Zone

When a negative gradient exists over a positive gradient in extremely deep water, a phenomenon known as a convergence zone may occur. Sound originating in the thermocline bends toward greater depths because of the decreasing temperature. However, as the sound penetrates into deep water, it passes the region of minimum sound speed and enters the deep isothermal layer. There the sound speed gradient operates in the other direction: the sound bends upward rather than downward and returns to the surface; see figure 9-20.

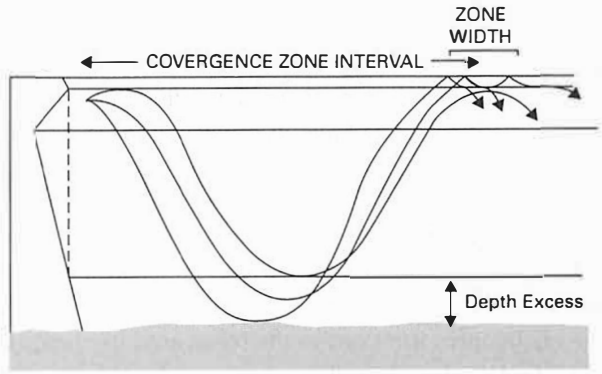


FIGURE 9-20 *Convergence zone propagation*

This produces a convergence zone, where the sound waves concentrate as if they had been focused. It typically lies at a distance of about 50 kilometers from the source. Beyond this convergence zone is a second zone of silence, where again the acoustic waves diffract downward; then another convergence zone, 50 kilometers out; and so forth.

The mapping of these zones is a routine part of submarine operations; by measuring deep-water temperatures with a bathythermograph, data are obtained that readily allow a computer to calculate the appropriate sound paths.

The convergence zone path is shown as having depth based on the concept of a wide sound beam or "bundle" of sound rays emanating from a source. By definition, the convergence zone is formed when the upper ray of the bundle becomes horizontal. This is the depth at which the sound velocity is equal to the largest of either the velocity at the surface or the velocity at the bottom of the surface layer. The difference in depth between the ocean bottom and the depth at which the upper ray becomes horizontal is called the depth excess.

The depth excess defines the depth of the bundle of rays or "amount of sound" that forms the convergence zone path. As the sound energy travels upward and

approaches the surface, the path narrows, which tends to focus the sound energy and results in convergence gains. As a general rule of thumb, a minimum of a 400-meter depth excess is required to produce an operationally useful convergence zone with reliable convergence gains.

Bottom Bounce

In addition to being refracted by varying conditions in the medium, sound can be reflected in the manner of a light beam striking a mirrorlike surface and lose little of its intensity. The two surfaces that can produce this type of reflection are the surface of water and the bottom of the ocean. Rarely, if ever, are these surfaces smooth enough to give a mirrorlike reflection, but in many instances the majority of the sound is reflected as a beam. Some sonars make use of this phenomenon, which is called bottom bounce.

A beam is directed at an angle greater than 25° to ensure layer penetration toward the bottom, from which the beam is then reflected back to the surface. From the surface it is reflected back to the bottom again; see figure 9-21. Thus, the ray bounces from one to the other until its energy is dissipated or until it strikes a target and returns to the sonar. As with reflected light, the angle of reflection is equal to the angle of incidence; therefore, ranging over a fairly flat bottom is more effective than ranging over a rough or sloping bottom. Additionally, the bottom bounce path is highly dependent upon depth and absorption of sound by the ocean bottom.

Multiple Paths

It is obvious, then, that many sound paths are available, as indicated in the composite drawing in figure 9-22. For simplicity in the figure, the bottom bounce path is depicted as a single beam, whereas in the real world, a wider beam corresponding to a multidirectional sound source is the type of beam used. Although many paths are available, only by observing the environment carefully and paying close attention to the equipment will the operator be able to use the various paths to the best advantage and not find them a liability.

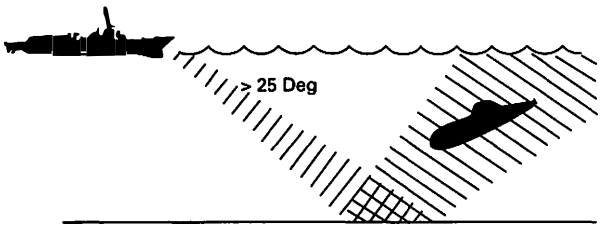


FIGURE 9-21 *Bottom bounce propagation*

Environmental Problems

As has been previously stated, the environment's variability contributes to the degree of difficulty of the under-water detection problem. As a fortunate counter to its variability, the speed profile of the environment is measurable, but it is measurable to both the hunter and the hunted. The understanding and exploitation of the acoustic environment determines who holds the detectability/counterdectability edge in the encounter. It is generally accepted that the advantage in undersea warfare belongs to the one who gains the initial detection.

Some of the environmental factors that will affect own-ship detectability by the enemy and own-ship capability for detection are listed below:

- Sound propagation paths available
- Shallow water vs. deep water
- Layer depth
- Seasonal variation in area (wind, temperature, etc.)
- Local transient phenomena (rain, sun warming upper layer, etc.)
- Currents in area of operations

Understanding the propagation paths available is paramount in assessing the counterdetectability situation. This understanding is based on past marine geophysical survey data, correlated with current bathymetry or sound velocity profile measurements. The viability of the bottom bounce propagation path is determined by survey data on bottom reflectivity and bottom absorption loss. The existence of convergence zones is normally based on ocean depth and knowledge of prior surveys. In the absence of prior surveys, ocean depth is an acceptable basis for predicting the absence or presence of a convergence zone.

The depth of the surface layer is a great determinant of hull-mounted surface-ship sonar performance because the target submarine may operate below the layer. As was discussed earlier, cross-layer detection is usually limited in range because of the refraction or bending of the sound rays. The tactical solution to this situation for the surface ship may be to vary the vertical angle of transmission of the sonar projector so as to penetrate the layer, or to deploy a variable-depth sensor below the layer.

Predictions of seasonal variation in the area of operations, based on previous surveys and observations of wind action and temperature profiles, provide initial data for the operational planner. Prior to arrival in the area of

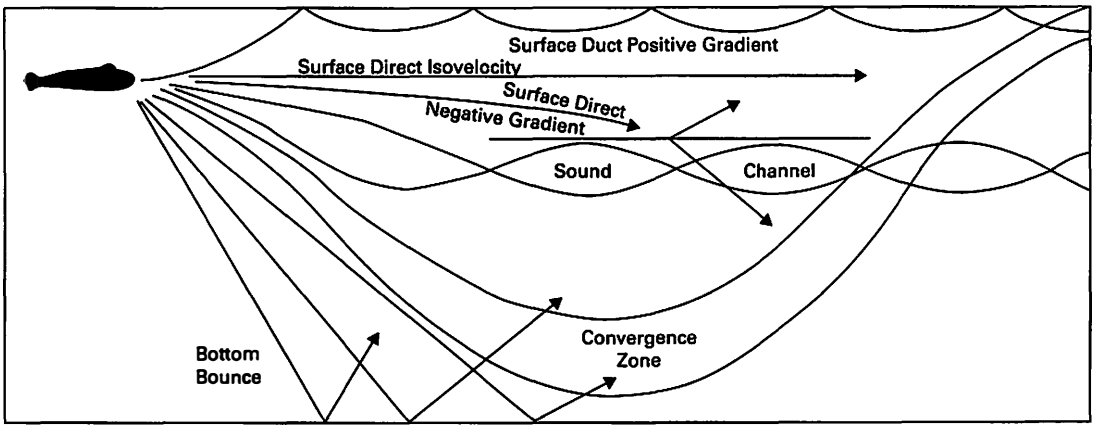


FIGURE 9-22 *Composite multipath representation*

operations, an estimate of the acoustic environment may be made using historical data and data obtained and transmitted by other platforms.

Upon arrival in the area of operations, the tactical commander should validate or verify the predictions through regular, periodic measurements of the ocean temperature or sound speed structure. The length of the period between measurements should be based on observations of current weather phenomena affecting sound propagation conditions, such as wind, time of day, and so on.

It is particularly important to be aware of the surface-heating effect of the sun and of the fact that in areas of little mixing of the sea by wind-driven waves, a positive temperature gradient may develop between midmorning and midafternoon that could seriously degrade surface-ship sonar performance. Under these conditions, frequent environmental measurements are in order.

Similarly, when conducting operations in the vicinity of currents like the Gulf Stream or the Labrador Current, where "fingers" of water with marked differences in temperature from that of adjacent waters is common, the tactical commander should consider more frequent measurements.

Continental shelf regions where water depths are generally less than 100 fathoms are known as "shallow water" regions. Many tactically significant areas that tend to concentrate maritime traffic—such as straits, choke points, and harbor approaches—lie in shallow water. Additionally, the end of the Cold War brought a shift in focus from "blue water" to littoral naval operations, which dramatically increased the proportion of operations conducted in shallow water.

Shallow water regions are particularly complicated and challenging acoustic environments for the following reasons:

1. The close proximity of the ocean surface to the bottom produces complex sound paths involving numerous reflections between the two surfaces and increases sound scattering and reverberation (backscattered noise). This close proximity also minimizes long-range propagation paths, tending to shorten acoustic detection ranges compared to the open-ocean case.
2. Noise levels are generally higher due to increased ship traffic density, increased concentrations of marine life, and increased operations such as fishing or offshore drilling in shallow water areas.
3. Highly variable water temperatures and salinities due to river outflows and tidal effects lead to highly variable and unpredictable sound paths. Irregular bathymetry from bottom debris and pinnacles further complicates the situation, as do highly variable bottom materials.
4. The bottom is likely to be sloped, resulting in changes in depth with distance ("upslope" or "downslope" sound propagation). Depending on the situation, such effects can either increase or decrease acoustic detection ranges.

Figure 9-23 is an example of the extremely complicated acoustic conditions found in shallow water regions. The complex nature of shallow water acoustics requires increased monitoring of the acoustic environment and increased attention given to predictive methods.

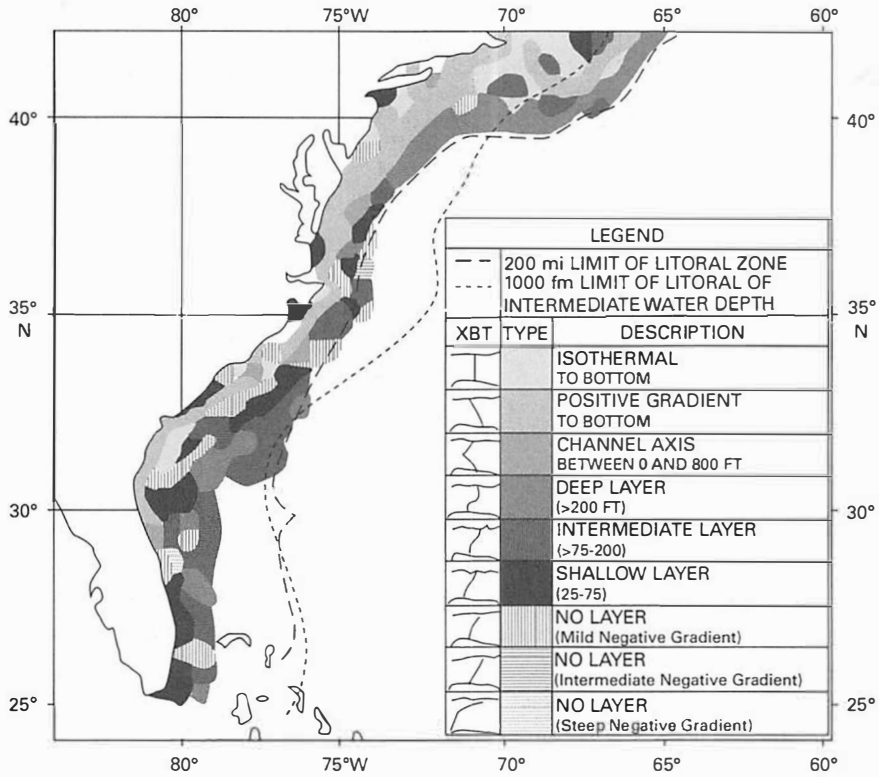


FIGURE 9-23 *Acoustic variability along the east coast of the United States*

Sound Sources and Noise

Background noise interferes with the reception of acoustic signals of interest. Noise is produced by a variety of sources, which can be classified into two general categories: self-noise and ambient noise. Self-noise, abbreviated NL_{self}, is associated with the electronic and mechanical operation of the sonar and the platform. Ambient noise, abbreviated $NL_{ambient}$, encompasses all external noise sources.

Self-Noise

Self-noise is produced by components in the sonar circuitry, water turbulence around the housing of the transducer, loose structural parts of the hull, machinery, cavitation, and hydrodynamic noises caused by the motion of the ship through the water. Self-noise sources, as well as all other acoustic sources, fall into two main categories: broadband and narrowband.[vi]

Broadband sources, as the name suggests, create acoustic energy over a wide range of frequencies; see figure 9-24. An analogy in optics is white light or thermal radiators, which are composed of a broadband of many different electromagnetic frequencies. Typical broadband sources are propellers/shafts, flow noise over the ship's hull, or the entire ship's propulsion system. Noise from the propeller and shaft is generally at a low frequency, meaning less than 1,000 Hz.[vii]

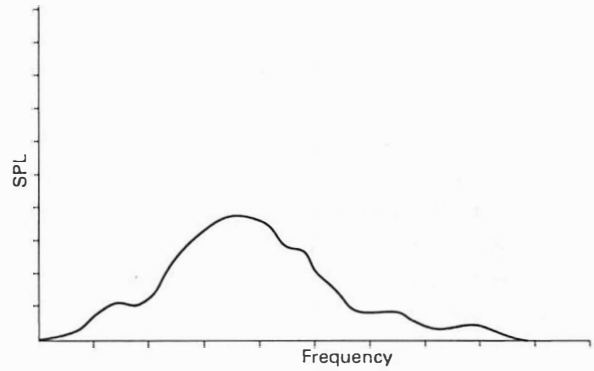


FIGURE 9-24 *Broadband source*

Narrowband sources radiate within a small band or class of frequencies; see figure 9-25. An optical analogy is a laser source or selective radiators, which contain essentially one frequency in a very narrow band. Typical sources are the various pieces of machinery found in every ship or submarine, such as pumps, motors, electrical generators, and propulsion engines. When specifying

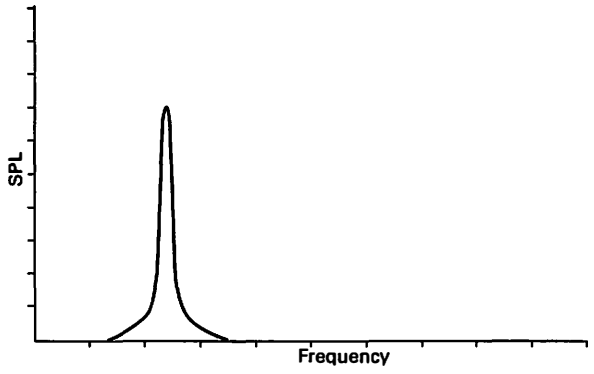


FIGURE 9-25 *Narrowband source*

narrowband sources, it is important to also specify the frequency at which they emit.[viii]

Machinery Noise

The dominant source of machinery noise in a ship is its power plant and the power distribution system that supplies power to the other machinery on the ship, such as compressors, generators, propellers, and so on. Machinery noise is normally always present and is kept to a minimum by acoustically isolating the various moving mechanical components.

The propulsion train is a large and important source of machinery noise. If the engine runs at a relatively low speed, as is the case with a reciprocating engine, gears may not be required between the engine and the propeller, making the engine the main noise source. High-speed engines such as steam or gas turbines, however, usually require reduction gears to connect them to the propeller.

The frequency of the explosions in the cylinders of a reciprocating engine is not likely to be a source of ultrasonic (above 15 KHz) noise but might be an important source of low-frequency sonic noise. A more crucial source of noise in the reciprocating engine is the clatter of the valves opening and closing. In gas turbines, the noise generated is in the ultrasonic region at a radian frequency equal to the angular velocity of the turbine buckets.

Noise in the ultrasonic region is extremely important in sonar and acoustic torpedo performance. The noise produced by auxiliary units such as pumps, generators, servos, and even relays is often more significant than the power plant noise. The large masses involved in the power plant usually keep noise frequencies relatively low. For this reason, a small relay may interfere more with the operation of a torpedo than an electric motor that generates several horsepower. Small, high-speed servomotors, however, may be serious sources of ultrasonic noise.

Flow Noise

Flow noise results when there is relative motion between an object and the water around it. This flow is easiest to understand by assuming that the object is stationary and that the water is moving past it. Under ideal conditions where the object is perfectly streamlined and smooth, water movement will be even and regular from the surface outward, as shown by the flow lines in figure 9-26A. This idealized condition is called laminar flow and produces no self-noise. Irregular objects can achieve nearly laminar flow conditions only at very low speeds—that is, below 2 knots (kts).

As flow speed increases, friction between the object and the water increases, resulting in turbulence—see figure 9-26B—and progressively increasing noise due to fluctuating static pressure in the water. If a hydrophone is placed in such a region, fluctuations of pressure will occur on its face, and it will measure the result in flow noise in the system.

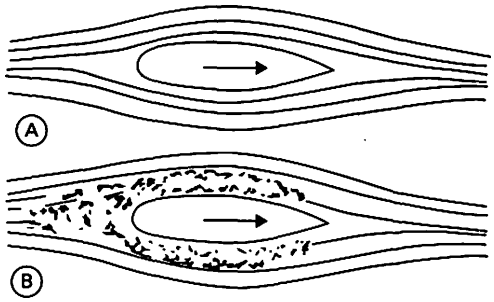


FIGURE 9-26 *Laminar versus turbulent flow*

As pressures fluctuate violently at any one point within the eddy, they also fluctuate violently from point to point inside the eddy. Moreover, at any given instant the average pressure of the eddy as a whole differs only slightly from the static pressure. Thus, very little noise is radiated outside the area of turbulence, and although a ship-mounted hydrophone may be in an intense flow noise field, another hydrophone at some distance from the ship may be unable to detect the noise at all. Flow noise, then, is almost exclusively a self-noise problem.

Actually, not much information is known about flow noise, but these general statements may be made:

1. It is a function of speed and sea state with a sharp threshold. At very low speeds, there is no observable flow noise. A slight increase in speed changes the flow pattern from laminar to turbulent and strong flow noise is observed immediately. Further increases in speed step up the intensity of the noise.

2. It is essentially a low-frequency noise.
3. It has very high levels within the area of turbulence, but low levels in the radiated field. In general, the noise field is strongest at the surface of the moving body and decreases rapidly with distance from the surface.
4. The amount of flow noise can be related to the degree of marine fouling on the ship's bottom and sonar dome. Figure 9-27 shows self-noise of a representative destroyer in sea state 2 as a function of speed at 10,000 Hz.

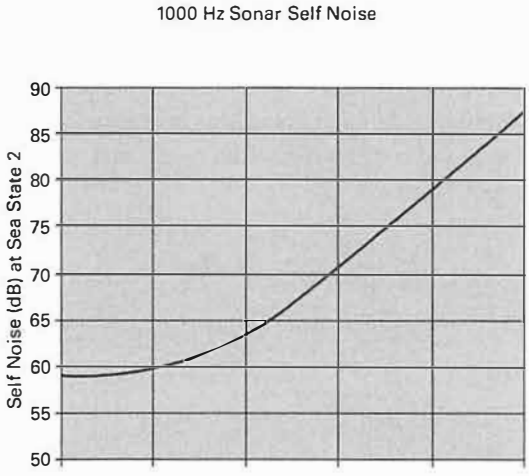


FIGURE 9-27 *Self-noise levels versus speed of a destroyer*

Fouling is the attachment and growth of marine animals and plants upon submerged objects. More than 2,000 species of organisms are known to foul, but only about 50 to 100 species present significant problems. The problem-causing organisms can be divided into two groups: those with shells and those without. Particularly important of the shell forms are tube worms, barnacles, and mollusks because they attach firmly and resist being torn loose by water action.

Nonshelled forms such as algae, hydroids, and tunicates commonly attach to relatively stationary objects and therefore do not cause much trouble on ships that frequently get underway.

Most fouling organisms reproduce by means of fertilization, and the resulting larvae drift with the currents. Depending on the species, the larval stage may vary from a few hours to several weeks. Unless the larvae attach themselves to a suitable surface during this period, they will not grow into adulthood.

Geography governs fouling by isolating species with natural barriers. Temperature is the most important factor governing distribution of individual species, limiting reproduction or killing adults. On a local scale, salinity, pollution, light, and water movement affect the composition and development of fouling communities. Most fouling organisms are sensitive to variations in salinity, developing best at normal sea water concentrations of 30 to 35 parts per thousand.

Pollution, depending on type, may promote or inhibit fouling. Fouling algae and microscopic plants, the food supply of many fouling animals, are dependent upon sufficient light for proper growth.

To forestall fouling on ship bottoms and sonar domes, the Navy uses special paints with antifouling ingredients. The most common active ingredient of antifouling paint is cuprous oxide, and the paint is effective from 2 to $2^{1}/_{2}$ years.

Cavitation

As the speed of ship or object is increased, the local pressure drops low enough at some points behind the object to allow the formation of steam. This decrease in pressure and the resulting bubbles of vapor represent the onset of cavitation. As the object moves away from the bubbles, however, the pressure increases, causing the bubbles to collapse and produce a continuous, sharp, hissing noise signal that is very audible.

Because the onset of cavitation is related to the speed of the object, it is logical that cavitation first appears at the tips of the propeller blades, inasmuch as the speed of the blade tips is considerably greater than that of the propeller hub. This phenomenon, known as blade-tip cavitation, is illustrated in figure 9-28.

As the propeller speed increases, a greater portion of the propeller's surface is moving fast enough to cause cavitation, and the cavitating area begins to move down the trailing edge of the blade. As the speed increases further, the entire front face of the blade commences cavitating, producing what is known as sheet cavitation, shown in figure 9-28.

The amplitude and frequency of cavitation noise are affected considerably by changing speeds (and changing depth in the case of a submarine). As speed increases, so does cavitation noise. As depth increases, the cavitation noise decreases and moves to the higher-frequency end of the spectrum almost as though speed had been decreased. This decrease in noise is caused by the increased pressure with depth, which inhibits the formation of cavitation bubbles.

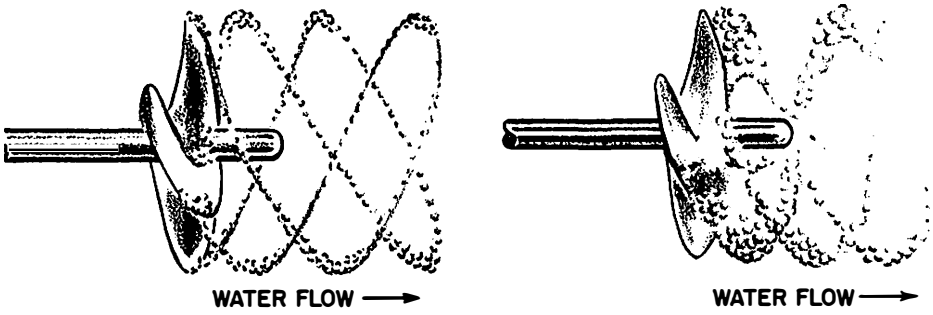


FIGURE 9-28 *Cavitation*

A given propeller at a given depth of submergence will produce no bubbles unless its speed exceeds a certain critical value. When the speed (in rpm) exceeds this critical value, the number of bubbles formed increases rapidly, but not according to any known law. The critical speed itself, however, depends simply on the depth of the propeller beneath the sea surface. This dependence is given by[ix]:

$$\frac{V_1^2}{V_2^2} = \frac{P_1}{P_2} \tag{9-20}$$

where:

V_1 = initial velocity when cavitation occurred at pressure P_1

V_2 = velocity when cavitation occurs at P_2

Thus, if a given propeller begins to cavitate at 50 rpm when at a depth of 15 feet, it will begin to cavitate at 100 rpm when at a depth of 60 feet.

Since all torpedo homing systems and many sonar systems operate in the ultrasonic region, cavitation noise is a serious problem. Torpedoes generally home on the cavitation noise produced by ships, and any cavitation noise produced by the torpedo interferes with the target noise the torpedo receives.

Because the speed at which a vehicle can operate without cavitating increases as the ambient pressure is increased, some acoustic torpedoes are designed to search and attack from depths known to be below the torpedoes' cavitating depth. In the same way, a submarine commander when under attack will attempt to dive to depths where it is possible to attain a relatively high speed without producing cavitation.

A simple mathematical model was developed for surface ships that relates cavitation sound pressure levels to frequency and ship speed.

$$NL_c = 190 + 53\log\left(\frac{v}{10\text{kts}}\right) - 20\log(f) \tag{9-21}$$

$f) 100\text{Hz}$

where:

v = ship's speed (kts)

f = frequency of interest (Hz)

A similar model would be expected to hold for submarines as well.

Quieting

It is important for ships engaged in undersea warfare to operate with a very low self-noise level. Noise reduction methods can be applied to other military ships, and for that matter to commercial ships. For protection against acoustic mines and reduction of the risk of being detected early by sonar listening devices, noise minimizing is necessary. A low level of noise also permits a ship to receive a lower level of signal from the enemy. Strict levels on the limits of structureborne noise and airborne noise also improve the habitability for the ship's crew.[x]

Surface Quieting

Consideration must be given when designing a ship to minimizing noise emission. Attention must be paid to the layout of the propellers and the hull's appendages with regard to waterborne noise generation. Model tests on propellers are necessary to check cavitation and to shift the cavitation point to a relatively high speed. Above this speed a propeller air emission system can be used to reduce cavitation-generated waterborne noise.

In the United States, this bubble-injection system is called Prairie Masker. "Prairie" is an acronym for *p*ropeller *air*-induced *e*mission. A system on some U.S. surface combatants for injecting air through hull girths is called Masker. This somewhat reduces the flow noise; see figures 9-29 and 9-30.[xi]

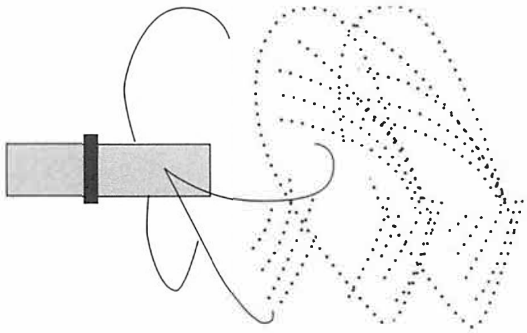


FIGURE 9-29 *Prairie air system*

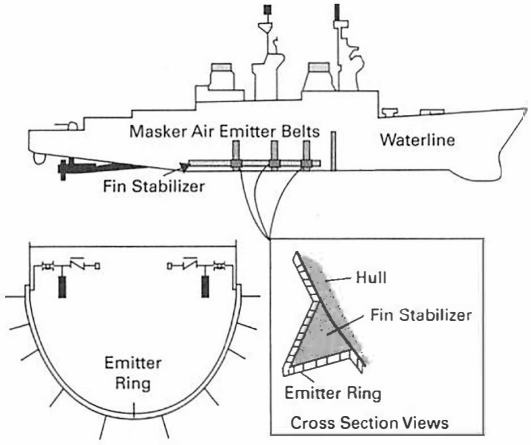


FIGURE 9-30 *Masker belts*

Other measures that can be taken to reduce noise are the following:

1. Reduce structureborne noise transmission through the use of flexible mounts for critical items combined with high impedance foundations; see figure 9-32.
2. Control noise in the transmission path—that is, decks and bulkheads. This can be accomplished through isolation using resilient mounts, distributed isolation material (DIM), and flexible piping connections; see figure 9-33.
3. Control noise at the source—the equipment. Examples are replacing noisy machinery with quieter machinery; using quiet valves for steam, hydraulics, and sea water; and using Prairie air to mask propeller cavitation.
4. Control noise at the radiating surface—that is, the hull. Coating materials for the hull and masker air are the two methods used to acoustically isolate the hull from the water.
5. Encapsulate all diesel prime movers, compressors, and gas turbines. Provide for high attenuation silencers in exhaust ducts of all diesel prime movers and gas turbines as well as for the combustion air inlets of cruising diesels and gas turbines.[xii]

Special attention must be paid to diesel units, compressors, and converters positioned near the ship's sonar. Because those are considered to be the most critical structureborne noise sources, they should be double elastic-mounted. In addition, the intermediate foundation frame structure of the diesel units should be damped with reinforced concrete blocks at the foundation springs.

The design of the main reduction gear units is important to noise reduction as well. The rotating parts and meshes are manufactured to the highest accuracy to minimize gear contact noise. In addition, the reduction gear assemblies are hard elastic-mounted. In the design of elastic mounts, attention must be paid to tuning with regard to the potential vibration-exciting sources of the ship and its equipment. The behavior of the mounts under shock and rough sea conditions also must be examined.

Submarine Quieting

Self-noise is a limiting factor in sonar performance, both from a counterdetection standpoint and the platform's ability to detect and track the target. Thus, there is a strong incentive toward building quieter submarines.

Significant improvement was realized in using a "bed plate" or acoustically isolated foundation for the gears, turbines, condensers, and turbo generators of the propulsion plant. Its noise-isolating mountings, however, added a good deal of volume and weight when compared with



FIGURE 9-31 *Combatant with Prairie Masker turned on (left)*

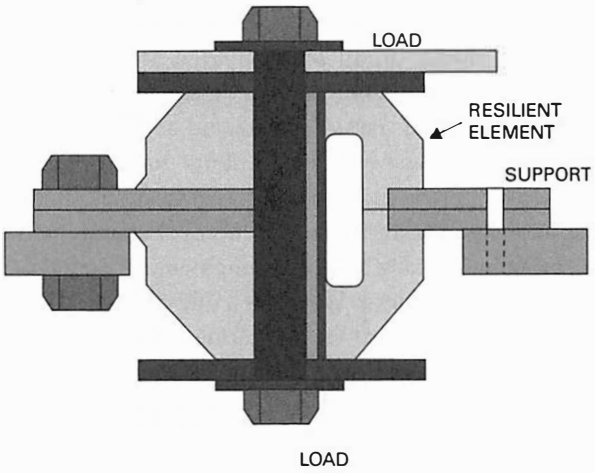


FIGURE 9-32 *Resilient mounts—rubber isolation used on machinery and piping*

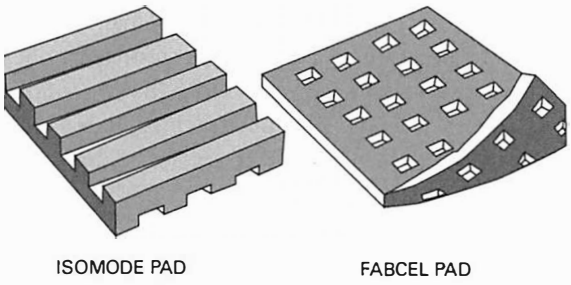


FIGURE 9-33 *Distributed isolation material (DIM)—rubber-type pads used on smaller equipment*

the conventional alternative of mounting the machinery directly to the hull.

Along with isolating noise sources inside the submarine, considerable progress has been made in reducing the noise levels generated by these sources. Rotating equipment such as motors, pumps, and fans are quieted by balancing the rotating components of this equipment, or by designing the submarine to have fewer pieces of rotating, reciprocating, or otherwise moving equipment.

The reduction gearing, which converts the power of the rapidly rotating turbine shaft to that of the slowly rotating propeller, is a particularly strong noise source. Past experimental noise reduction efforts included replacing the propulsion steam turbine and reduction gear with a large electrically driven motor to turn the propulsion shaft. Although this approach met with limited success at the time, recent advances in technology have led to new interest in electric propulsion.

Significant headway has also been made in reducing propeller cavitation. Further silencing gains have been made by using pump-jet propulsors and hull-mounted sound-absorbing acoustic tiles.[xiii]

Ambient Noise

Ambient noise is background noise in the sea that is due to either natural or man-made causes and may be divided into four general categories: hydrodynamic, seismic, ocean traffic, and biological.

Hydrodynamic Noise

Hydrodynamic noise is caused by the movement of the water itself as a result of tides, winds, currents, and storms. The level of hydrodynamic noise present in the sea is directly related to the condition of the sea surface. As the surface becomes agitated by wind or storms, the noise level rises, thus reducing detection capability. Very high hydrodynamic noise levels caused by severe storms in the general area of the ship can result in the complete loss of all signal reception.

Wave height is directly attributable to sustained wind speed. Table 9-2 relates the Beaufort Wind Scale to wave height and sea state.

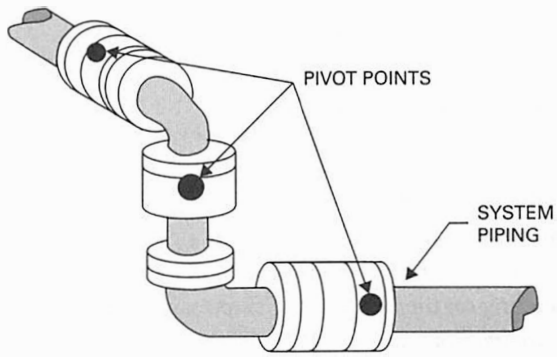


FIGURE 9-34 *Flexible connections used on pipes and hoses*

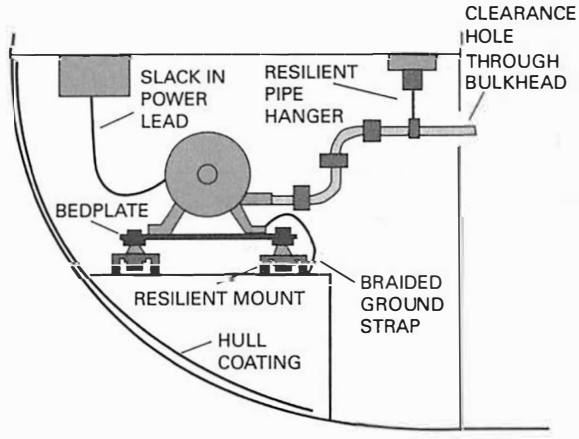


FIGURE 9-35 *Methods of reduction in self-noise*

TABLE 9-2 *Beaufort Wind Scale and sea states*

Beaufort	Wind Speed (knots)	Wave Height (ft)	Sea State
0	<1	0	0
1	1–3	0–0.25	
2	4–6	0.5–1	1
3	7–10	2–3	2
4	11–16	3.5–5	3
5	17–21	6–8	4
6	22–27	9.5–13	5
7	28–33	13.5–19	6
8	24–40	18–28	
9	41–47	23–32	
10	48–55	29–49	7
11	56–63	39–46	8
12	>64	37–52	9

Seismic Noise

Seismic noises are caused by land movements under or near the sea—as, for example, during an earthquake. They are fairly rare and of short duration, and hence will not be elaborated upon.

Ocean Traffic

The effect of ocean traffic on ambient noise level is determined by the area's sound propagation characteristics, the number of ships, and the distance of the shipping from the area. Noise caused by shipping is similar to that discussed regarding self-noise, with the frequencies depending on the range to the ships causing the noise.

Noises from nearby shipping can be heard over a wide spectrum of frequencies, but as the distance becomes greater, the range of frequencies becomes smaller, with only the lower frequencies reaching the receiver because the high frequencies are attenuated. In deep water, the low frequencies may be heard for thousands of kilometers. Table 9-3 shows the Wenz curve categories that will be used to determine ocean traffic noise.

TABLE 9-3 *Wenz ocean traffic categories*

Category	Description
Very remote	No other ships present for many miles
Remote	Infrequent distant ships
Quiet	Occasional ship nearby
Shipping lanes	Many ships nearby
Heavy traffic (deep water)	Constant passing of ships nearby, offshore
Heavy traffic (shallow water)	Constant passing of ships nearby, inshore

Biological Noise

Biological noises produced by marine life are part of ambient background noise and at times are an important factor in undersea warfare. Plants and animals that foul the ships are passive and contribute to self-noise by increasing water turbulence. Crustaceans, fish, and marine mammals are active producers of sounds, which are picked up readily by sonar equipment.

During and since World War II, a great deal of research on sound-producing marine animals had been carried out. The object was to learn all the species of animals that produced sound, their methods of production, and the physical characteristics of the sounds (their frequencies, intensities, etc.). Sounds produced by many species have been analyzed electronically, and considerable physical data have been obtained.

Important sonic marine animals are members of one of three groups: crustaceans, fish, or mammals.

Crustaceans, particularly snapping shrimp, are one of the most important groups of sonic marine animals. Snapping shrimp, about 2 centimeters long, bear a general resemblance to the commercial species of shrimp, but are distinguishable from them by one long, large claw that has a hard movable finger at the end. They produce sound by snapping the finger against the end of the claw.

Distribution of snapping shrimp appears to be governed by temperature, and they are found near land in a worldwide belt lying between latitudes 35°N and 40°S. In some places, such as along the coast of Europe, they range as far north and south as 52°. The largest colonies or beds of snapping shrimp occur at depths of less than 30 fathoms on the bottoms of coral, rock, and shell. There are exceptions, however. They may, for example, occur as deep as 250 fathoms, and they have been found on mud and sand bottoms covered with vegetation.

A shrimp bed is capable of producing an uninterrupted crackle resembling the sound of frying fat or burning underbrush. Frequencies range from less than 1 KHz to 50 KHz. Noise is constant, but there is a diurnal cycle, with the maximum level at sunset. Over beds, a pressure level of 86dB referenced to 1μPa has been noted. Intensity drops off rapidly as the range from the bed increases. Lobsters, crabs, and other crustaceans also make minor contributions to background noise.

Fish produce a variety of sounds that may be placed in three categories, depending upon how the sounds are made. The first category includes sound produced by the air bladder, a membranous sac of atmospheric gases lying in the abdomen. The sound is caused by the movement of muscles inside or outside the bladder or by the general movement of the body.

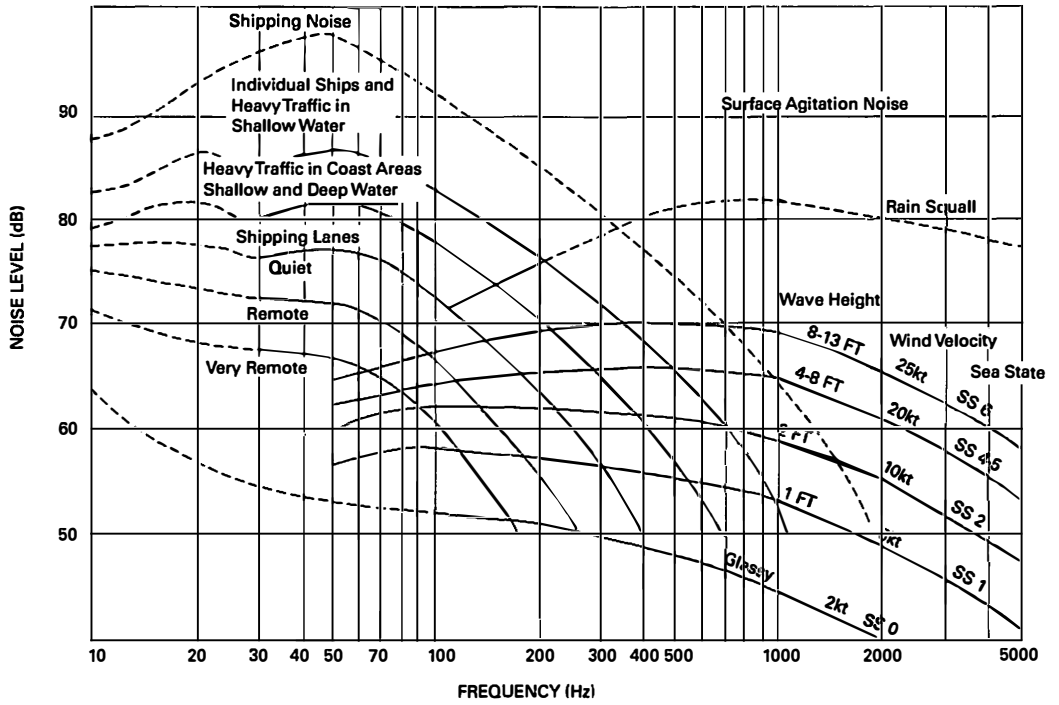


FIGURE 9-36 *Wenz ambient noise curves*

The second division includes sounds produced by various parts of the body such as fins, teeth, and the like, rubbing together. This noise is called stridulatory sound.

The third class includes sounds that are incidental to normal activities, such as colliding with other fish or the bottom when swimming, biting, and chewing while feeding, and so on.

The majority of sonic fish inhabit coastal waters, mostly in temperate and tropical climates. Although fish are the most prevalent—and therefore the most important—sound producers, their activity is not as continuous in intensity as that of snapping shrimp.

The level of sound produced by fish increases daily when they feed (usually at dawn and dusk) and annually when they breed. Fish sounds range in frequency from about 50 to 8,000 Hz. Sounds of air bladder origin have most of their energy concentrated at the lower end of this spectrum, 75 to 150 Hz, whereas stridulatory sounds characteristically are concentrated at the higher end of the spectrum.

Marine mammals, in addition to returning echoes from sonar equipment, produce sound vocally and stridulously. Seals, sea lions, and similar animals possess vocal cords and bark or whistle or expel air through their mouths and nostrils to produce hisses and snorts. Whales, porpoises, and dolphins force air through soft-walled nasal sacs, and produce sounds that have been described as echo-ranging pings, squeals, and long, drawn-out moans similar to a foghorn blast. Other sounds probably produced by stridulation and attributed principally to whales, porpoises, and dolphins are described as clicking and snapping.

Wenz Curve

Although difficult to determine accurately on an operational basis, ambient noise levels are nonetheless important factors to be considered in determining sonar performance. Figure 9-36 is an expansion of the shipping noise/wind noise portion of the Wenz curves. Figure 9-37 is an example of a set of Wenz curves that can be used to estimate noise levels from a variety of sources within the frequency range of interest.

Sample Ambient Noise Problem

A SOSUS station is searching for an enemy submarine known to be producing sound at a frequency of 300 Hz. It is suspected that the sub is patrolling in shipping lanes that currently have 6-foot seas. Use Wenz curves to determine an approximate value of ambient noise.

From figure 9-37, the ambient noise level due to shipping at 300Hz is 65dB and the noise level due to 6-foot seas is 66dB.

Using the nomogram in figure 9-3, combine the signals of 65dB and 66dB.

$$IL_1 - IL_2 = 66 - 65 = 1$$

From the nomogram, add 2.4dB to 66dB. This equals 68.4dB.

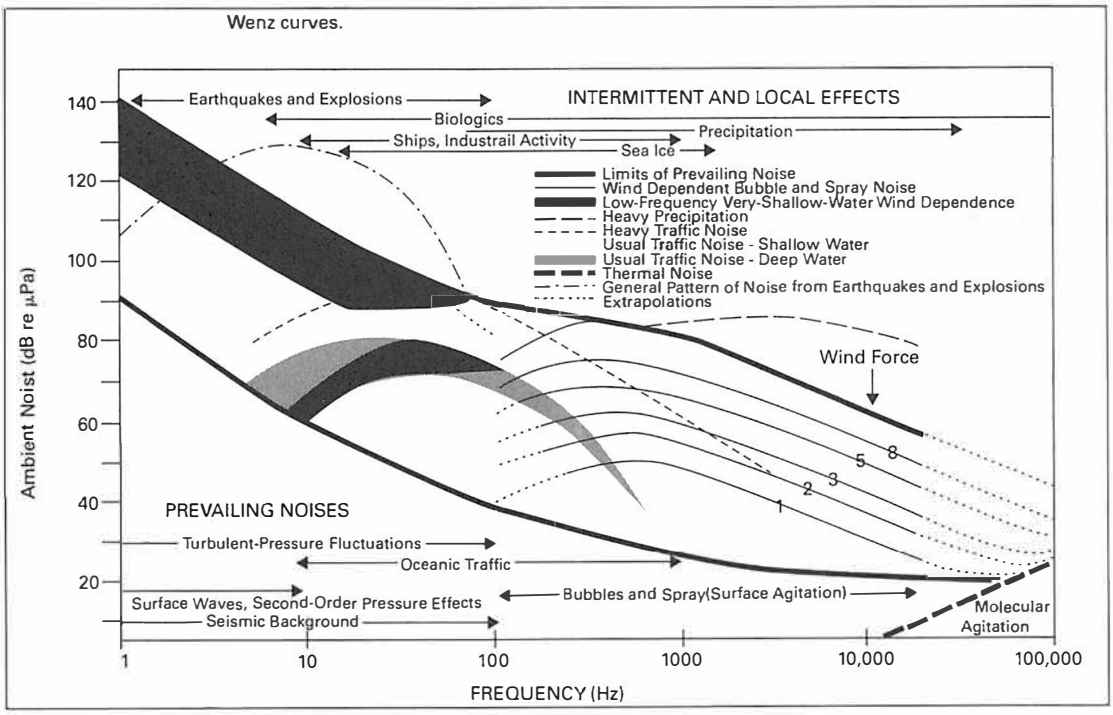


FIGURE 9-37 *Wenz acoustic ambient noise in the ocean curves*

---

[i] Hall, 178

[ii] "Fundamentals," sec. 8.3.3

[iii] Urick, 113

[iv] "Fundamentals," sec. 8.4.7

[v] Urick, 164

[vi] Hall, 194

[vii] Hall, 194

[viii] Hall, 195

[ix] "Fundamentals," sec. 8.8.1.3

[x] "Fundamentals," sec. 8.8.1.4

[xi] "Ship's"

[xii] "Ship's"

[xiii] "Fundamentals," sec. 8.8.1.4.2

10 Figure of Merit

Introduction

The key to success in undersea warfare is detecting an adversary before being detected. The major sensor in use today is sonar, and the present state of knowledge of the physical world suggests that sound will be used for many years to come. An understanding of sonar is achieved through the comprehension of the sonar equations and a concept called figure of merit (FOM).

Many phenomena and effects associated with underwater sound are conveniently and logically related in a quantitative manner by the sonar equations. For many problems in undersea warfare, the sonar equations are the working relationships that tie together the effects of the medium, the target, and the equipment so that an operator can effectively use and understand the information received and provide additional information such as range prediction.

Therefore, the purpose of this chapter is to spell out the sonar equations and the figure of merit, to state the specifics of their usefulness, and to indicate how the various parameters in the sonar equations, including the figure of merit, can be measured.

The sonar equations are based on a relationship or ratio that exists between the desired and undesired portions of the received energy when some function of the sonar equipment such as detection or classification is performed. These functions all involve the reception of acoustic energy occurring in a natural acoustic background. Of the total acoustic energy at the receiver, a portion is from the target and is called the signal. The remainder is from the environment and is called noise.

As discussed in the previous chapter, oceans are filled with noise sources such as breaking waves, marine organisms, surf, and distant shipping, which combine to produce what is known as ambient noise. Self-noise, in contrast, is produced by machinery within the receiving platform and by motion of the receiving platform through the water. Further, in active systems, scatterers such as fish, bubbles, sea surface, and sea bottom produce an unwanted return called reverberation, which contributes to the masking of the desired signal.

The function of the design engineer is to optimize the signal-to-noise (SNR) ratio for all conditions for the sonar equipment. The operator then can predict the detection probability by using knowledge of these design specifications, the operator's own ability and training, and the predicted environmental conditions extrapolated from previously determined or actual onboard measurements.

Sonar Equation Terms

Detection Threshold

In order to predict performance, the acoustic problem must be defined or quantified in a manner that provides a measure of predictability for varying signal and noise levels. This quantity is known as the detection threshold (DT), and it attempts to describe in a single number everything that happens once the signal and its accompanying noise are received at the sonar. The detection threshold is defined as "the signal minus the noise level required inboard of the hydrophone array such that an operator can detect a target."

Actually, the task of detecting a sonar signal is a chance process for several reasons, one of which is that a human operator is involved. Designating a sonar contact as a target when a target is indeed present is defined as a "detection." To do so when a target is not present is defined as a "false alarm." The definition of DT therefore is normally qualified by adding the requirement that an operator "can detect a target using a particular piece of equipment on 50% of those occasions for which a target is present." Thus, if the average signal-to-noise value

provided equals the average of required signal-to-noise, detection occurs in 50% of the times that detection could occur. To summarize:

If average provided = average required, then detection probability is 50%.
If average provided > average required, then detection probability is 50% to 100%.
If average provided < average required, then detection probability is 50% to 0%.

Note: The instantaneous value of the provided or required signal-to-noise can vary over a wide range due to the variability of operators as well as time fluctuations in propagation loss, target radiated signal, and own-ship noise. The consequence of this is that although the average value of signal-to-noise provided may be less than the average value required, at times the instantaneous value may be greater than that required and detection may occur; accordingly, a probability of detection greater than zero always exists.

Putting this all together, it can be seen that if detection is to occur with a specified degree of probability, then the signal (expressed in decibels) minus the noise (expressed in decibels) must be equal to or greater than a number—the detection threshold—which also is expressed in decibels.

$$S - N \geq DT \qquad (10\text{-}1)$$

The next step is to expand this basic sonar equation relating signal-to-noise in parameters determined by the equipment, the environment, and the target.

Parameters determined by the equipment:

- Own active sonar source level: SL
- Self-noise level: NL_{self}
- Receiving directivity index: DI
- Detection threshold: DT

Parameters determined by the environment:

- Transmission loss: TL
- Reverberation level: RL
- Ambient noise level: $NL_{ambient}$

Parameters determined by the target:

- Target strength: TS
- Target passive source level: SL

Note that two pairs of the parameters are given the same symbol (own active sonar source level/target source level, and self-noise level/ambient noise level) because they are accounted for similarly in the sonar equations. This set of parameters is not unique, nor are the symbols the same in all publications, but they are the ones conventionally used in technical literature. It should be carefully observed as the discussion progresses how each of these parameters fits into the basic sonar equations.

Passive Source Level

The passive source level (SL) refers to the acoustic signal that is radiated by the target itself. The target signal can be caused by operating machinery, propeller noise, hull flow noise, and so on. The target's SL is, in essence its self-noise.

Of course, the source level is generally not known beforehand. However, source levels for hostile targets can be measured through intelligence-gathering efforts and tabulated for future reference. Passive sources fall into the two main categories of noise discussed in chapter 9, namely, broadband and narrowband.

Active Source Level

In an active sonar system, the source of the acoustic wave is the sonar system. Electrical energy from the transmitter is converted into acoustic energy by the sonar transducer. The output of the transducer is measured as an SPL and is also termed the source level (SL). For a given total power output, the intensity generally diminishes with increasing range because of spreading, absorption, and scattering. Therefore, the SPL, which is the logarithm of intensity, will decrease with range from the source.

We must measure SPL at a standard range from the source in order to be able to make meaningful comparisons. That standard range is 1 meter. Therefore, SL is specified by an SPL measured at 1 meter from the source. Sometimes the notation will include a reference to the pressure unit and measurement distance. For example,

$$SL = 145\text{dB}$$

$$\text{ref: } 1\ \mu\text{Pa @ } 1\ \text{m}$$

refers to a source level using a reference of 1 μPa measured at 1 meter. Here, we will omit this cumbersome notation since all source levels will be specified using the same reference.[i]

Transmission Loss

The ocean is a remarkably complex medium for the propagation of sound, as introduced in chapter 9. It possesses an internal structure and upper and lower boundaries that create many diverse effects upon the sound emitted from an underwater source. In traveling through the sea, an underwater sound signal is delayed, distorted, and weakened. The transmission loss term in the sonar equations expresses the magnitude of these effects.

Consider a source of sound located in the ocean. The intensity of the sound can be measured at any point in the ocean and at any range from the source. For purposes of measuring intensity at the source, the intensity measurement is generally taken at one unit distance from the source and labeled I_0. The intensity can then be measured at any distant point where a hydrophone is located and denoted I.

It is operationally significant to compare the two values. One way to do this is to form the ratio I_0/I. Note that if the ratio, denoted n, is greater than unity, the intensity at the source is greater than at the receiver, as would be expected. If $n = I_0/I$, then

$$10\log n = 10\log I_0 - 10\log I \qquad (10\text{-}2)$$

= sound intensity level at the source minus sound intensity at the receiver

The value 10 log n is called the transmission loss and is measured in decibels. Most of the factors that influence transmission loss have been accounted for by scientific research and can be grouped into categories primarily consisting of spreading, absorption, and scattering.

Spreading Loss

To understand spreading loss, first imagine a theoretical ocean that has no boundaries and in which every point has the same physical properties as any other point—that is, an infinite, homogeneous medium. In such a medium, sound energy would propagate from a point source radially in all directions along straight paths and would have a spherical wave front.

Under these conditions, the change in power density with distance from the point source would be due only to the spherical divergence of energy. Note that there is not a true loss of energy as might be implied; rather, the energy is simply spread over a progressively larger surface area, thus reducing its power density. For this model, the amount of power spread over the surface of a sphere of radius r, centered at a point source, is expressed by:

$$\text{power density} = \frac{P_t\,(\text{watts})}{4\pi r^2} \qquad (10\text{-}3)$$

where:

P_t = acoustic power level immediately adjacent to the source

r = radius from point source

This concept of power density (watts/m^2) was used to develop the radar equation in chapter 3. As stated here, the units of acoustic energy are watts/m^2. Therefore, equation 10-3 can be written:

$$I_r = \frac{P_t}{4\pi r^2} = \frac{P_e^2}{\rho c} \qquad (10\text{-}4)$$

where:

P_e = RMS pressure measured at range r

The intensity of the sound immediately adjacent to the source is measured, by convention, at a unit distance (1 meter) from the source. It will be labeled I_1 and is given by

$$I_1 = \frac{P_1}{4\pi(1)^2} \qquad (10\text{-}5)$$

The acoustic intensity (I_r) at some distance r from the source is less than the acoustic intensity (I_1) at 1 meter from the source. This is the result of spreading a fixed amount of power over a geometrically increasing surface area of a sphere; see figure 10-1. The reduction of the acoustic intensity as a function of distance r is shown in the ratio:

$$\frac{I_r}{I_1} = \frac{\frac{P_t}{4\pi(r)^2}}{\frac{P_t}{4\pi(1)^2}} = \frac{1}{r^2} \qquad (10\text{-}6)$$

However, the ratio I_r to I_1 at any significant range is typically so small that these values are best examined on a logarithmic scale using the decibel.

$$\begin{aligned} 10\log\frac{I_r}{I_1} &= 10\log\frac{1}{r^2} \\ &= 10\log(1) - 10\log r^2 \\ &= -20\log r \end{aligned} \qquad (10\text{-}7)$$

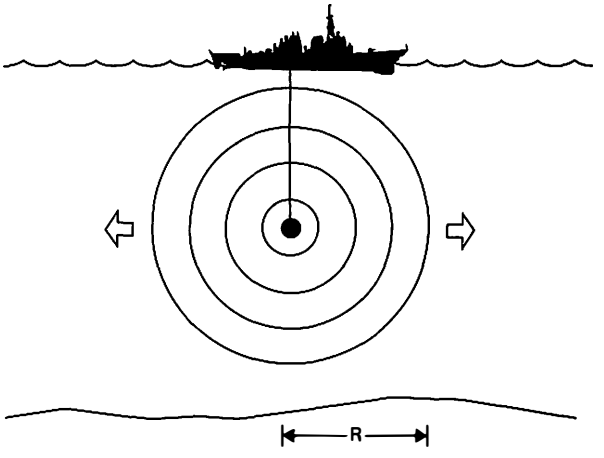


FIGURE 10-1 *Spherical spreading*

The reduction in acoustic intensity (I_1) due to spreading is called the transmission loss (TL), and for spherical spreading the component is:

$$TL = 20 \log r \tag{10-8}$$

and

$$10 \log I_r = 10 \log I_1 - TL \tag{10-9}$$

However, the ocean is not an unbounded medium and not all sources are omnidirectional point sources as assumed. For sources that radiate energy only in a horizontal direction, sound energy diverges more like the surface of an expanding cylinder. Also, since the ocean is bounded at the surface and at the bottom, cylindrical divergence is usually assumed for ranges that are large compared to the depth of the water or when sound energy is trapped within a thermal layer or sound channel. For this model, the acoustic intensity of energy at the surface of a cylinder of radius r is expressed by:

$$I_r = \frac{P_t}{2\pi r h} \tag{10-10}$$

where:

h = vertical distance between upper and lower boundaries; see figure 10-2

The transmission loss is therefore:

$$10 \log \frac{I_r}{I_1} = 10 \log \frac{1}{r} = 10 \log(1) - 10 \log r = -10 \log r$$

or

$$10 \log I_r = 10 \log I_1 - 10 \log r$$

or

$$TL = 10 \log r \tag{10-11}$$

and

$$10 \log I_r = 10 \log I_1 - TL$$

Equation 10-11 represents cylindrical divergence of sound energy, sometimes referred to as inverse first-power spreading.

A problem with the spherical and cylindrical transmission loss models occurs when deciding where to predict the range that spreading transmissions from the spherical case change to the cylindrical. Many methods have been proposed to predict this "transition range," all of which are based upon the geometrical arrangement of the depth of the layer, the depth of the source, and the sound speed gradient. An example of a transition range approximation for the mixed layer bounded by the sonic layer depth and the sea surface is[ii]:

$$r_t = \sqrt{\frac{RH}{8}} = \sqrt{\frac{H}{H-d}} \tag{10-12}$$

where:

H = mixed layer thickness

d = depth of the source

R = radius of curvature = $c_n/g\cos\theta_n$ (where: c_n = speed of sound; g = gradient; θ_n = angle of ray leaving the source)

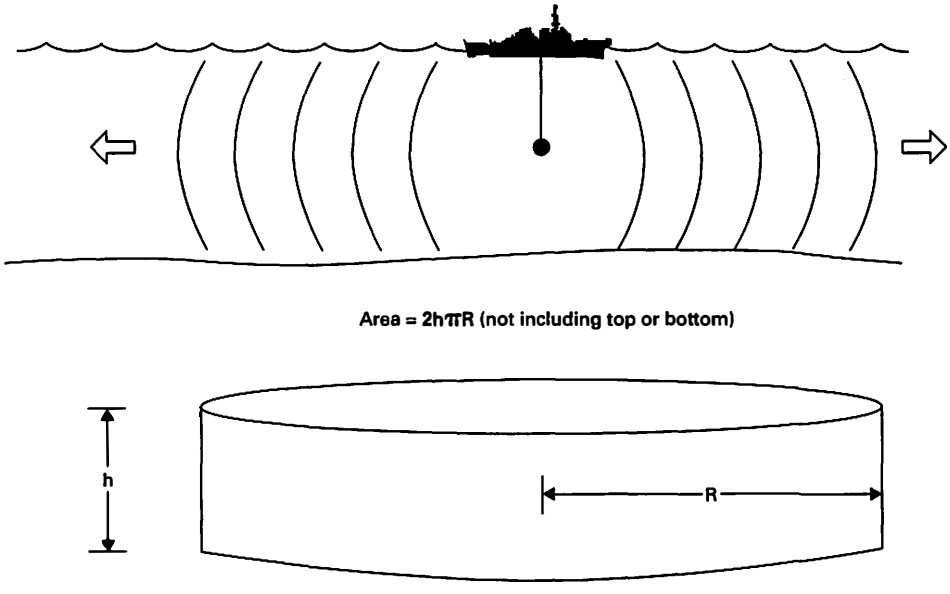


FIGURE 10-2 *Cylindrical spreading*

Equation 10-12 has severe limitations in that it is good only in the mixed layer; the radius of curvature is dependent upon a constant gradient; and it is good for only one angle from the source.

Because the transition range is difficult to find through a simple formula or model, the prediction is best determined by a complex computer model of the ocean with the source and receiver at given depths.

This text assumes that the transition range is 1,000 meters. This approximation is based on the average depth of the ocean, 2,000 meters, and the assumption that the source is located at 1,000 meters. Therefore the two equations for the spreading portion of transmission loss are:

$$TL = 20 \log r \qquad r < 1{,}000\,m \qquad (10\text{-}13)$$

$$TL = 10 \log r + 30 \qquad r \geq 1{,}000\,m \qquad (10\text{-}14)$$

Absorption Losses

Attenuation of sound energy in sea water arises principally through the action of two independent factors, absorption and scattering, with an additional contribution from bottom loss.

The primary causes of absorption have been attributed to several processes including viscosity, thermal conductivity, and chemical reactions involving ions in sea water, as summarized below:

1. Viscosity of the medium causes sound energy to be converted into heat by internal friction.
2. Some sound energy is converted into heat because sound waves alternately raise and lower the temperatures.
3. Suspended particles are set to oscillating by sound waves, and in this process some of the sound energy is dissipated in the form of heat. This is especially the case if the particles are air bubbles.

Although each of these factors offers its own unique contribution to the total absorption loss, all of them are caused by the repeated pressure fluctuations in the medium as sound waves are propagated. All the absorptive loss involve a process of converting acoustic energy into heat and thereby represent a true loss of acoustic energy to the environment.[iii]

Experimentation has produced a factor, called the absorption coefficient, that when multiplied by the range gives the total loss in dB due to absorption.

Water temperature and the magnesium sulphate ($MgSO_4$) concentration are important factors influencing the magnitude of the absorption coefficient value, because the colder the average water temperature and the greater the amount of $MgSO_4$ present, the greater will be the losses due to absorption.

However, it is the frequency of the sound wave that causes the most significant variation in the absorption coefficient. Although the formula will change slightly with temperature and geographical location, the general equation for the value of the absorption coefficient in decibels per meter for sea water at 5°C is:

$$\alpha = \frac{0.036 f^2}{f^2 + 3{,}600} + 3.2 \times 10^{-7} f^2 \qquad (10\text{-}15)$$

where:

α = absorption coefficient (dB/meter)

f = frequency in KHz

Even though this formula is rather cumbersome, the important thing to observe is that the absorption coefficient increases roughly as the square of the frequency. This relationship is of major importance to the naval tactician. It implies that if higher frequencies are chosen for sonar operation in order to achieve greater target definition, the price paid will be greater attenuation. In summary, the higher the frequency, the greater the attenuation, which causes shorter detection ranges.

For this reason, where long-range operation of sonar equipment is desired, the lower the frequency used the better. Figure 10-3 depicts typical values of the absorption coefficient of sea water at 5°C for varying frequencies. To obtain the transmission loss due to absorption, merely multiply the absorption coefficient by the range in meters. Thus:

$$TL = \alpha r \qquad (10\text{-}16)$$

Scattering Losses

Another form of attenuation is scattering, which results when sound strikes foreign bodies in the water and is reflected. Reflectors consists of boundaries such as the surface, bottom, and shores; other reflectors are bubbles, suspended solid and organic particles, marine life, and minor inhomogeneities in the thermal structure of the ocean.

The amount of energy scattered is also a function of the size, density, and concentration of foreign bodies present in the sound path, as well as the frequency of the sound wave. The larger the area of the reflector compared to the sound wavelength, the more effective the

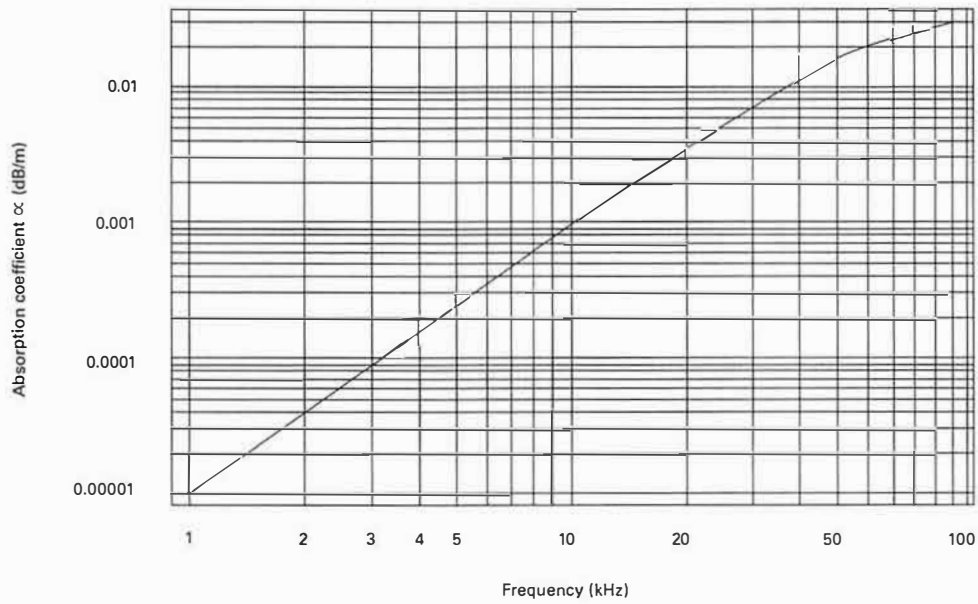


FIGURE 10-3 *Absorption versus frequency*

reflector is as a scatterer. Part of the reflected sound is returned to the source as an echo—is backscattered—and the remainder is reflected off in another direction and is lost energy. Backscattered energy is known as reverberation and is divided into three types: volume, surface, and bottom.

Volume reverberation is caused by various reflectors, but fish and other marine organisms are the major contributors. Additional causes are suspended solids, bubbles, and water masses of markedly different temperatures. Volume reverberation is always present during active sonar operations but is not normally a serious factor in masking target echoes.

The one exception involves the deep scattering layer (DSL), which is a relatively dense layer of marine life present in most areas of the ocean. During daylight hours, this layer is generally located at depths of about 600 meters and does not pose a serious problem. At night, however, the layer migrates toward the surface and becomes a major source of reverberation. This layer is rarely opaque to sound when detected with a sonar looking down on it from directly above, as with a fathometer, but this is not the case with a search sonar transmitting in a more or less horizontal direction. By transmitting horizontally, the sound waves encounter many more organisms, and the effect can vary from partial transmission of sound to total reflection and scattering, thereby hiding a submarine.

Surface reverberation is generated when transmitted sound rays strike the surface of the ocean—that is, the underside of the waves. It is always a factor in active sonar operations, and is directly related to wind speed because it controls wave size and the angle of incidence.

Bottom reverberation occurs whenever a sound pulse strikes the ocean bottom. In deep water this condition normally does not cause serious problems, but in shallow water, bottom reverberation can dominate the background and completely mask a nearby target's signal. The amount of energy lost through scattering will vary with the roughness of the bottom and the frequency of the incident sound.

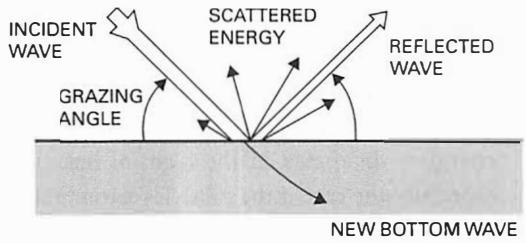


FIGURE 10-4 *Bottom absorption and scattering*

Sound reflected from the ocean floor usually suffers a significant loss in intensity. Part of this loss is caused by the scattering effects just described, but most of it results from the fact that a portion of sound energy will enter the bottom and travel within it as a new wave, as illustrated in figure 10-4. The net result is that the strength of the reflected wave is greatly reduced.

The amount of energy lost into the bottom varies with the bottom composition, sound frequency, and the striking

angle of the sound wave. The total of these losses can vary from as low as 2dB/bounce to greater than 30dB/bounce. In general, bottom loss will tend to increase with frequency and with the angle of incidence. Soft bottoms such as mud are usually associated with high bottom losses (10 to 30dB/bounce); hard bottoms such as smooth rock or sand produce lower losses.

Although it is possible to derive equations that will compute precise values of transmission loss due to volume, surface, and bottom loss factors, the ocean's characteristics are so variable that there is little utility in doing so. Therefore, it is customary in operational situations to make an educated guess of the scatter values and lump them together into one term called the transmission loss anomaly, symbolized A, which is included in the transmission loss equation.

Total Propagation Loss

It would be useful to have a simple mathematical relationship that would describe all the effects of the various factors influencing transmission loss as they occur in the ocean. But the state of the physical conditions encountered in the ocean is very complex and not at all easy to represent. A few mathematical models do exist that provide close approximations for some sets of conditions, but at present, no single model accounts for all the conditions encountered.

A simplified model used to obtain approximate values of transmission loss for the spherical spreading case for range <1,000 m is:

$$TL = 20\log r + \alpha r + A \qquad (10\text{-}17)$$

and for the cylindrical spreading case:

$$TL = 10\log r + 30 + \alpha r + A \qquad (10\text{-}18)$$

where range ≥1,000 m and "30" accounts for the spherical spreading in the first 1,000 meters.

It is important to realize that sound transmission in the ocean is three-dimensional and that transmission loss versus horizontal range alone is not sufficient information for most operational situations. Areas of no sonar coverage occur at various intervals of range because of refraction, reflection, and interference between waves traveling by means of different propagation paths.

Therefore, although the TL equations are interesting and somewhat useful, they are not always totally accurate. Often a more useful accounting for losses of sound through water can be obtained by using the term propagation loss, which is discussed later in the chapter.

Directivity Index

Directivity of an array can be expressed in terms of beamwidth usually measured at the half-power points of the beam. However, the term *directivity factor*, K, is more useful as a quantitative measure of array performance because it applies regardless of the shape of the beam and its sidelobes.

The directivity factor is appropriate for both line and curved (two-dimensional) arrays. The directivity factor is a direct corollary to the term of gain found in electromagnetic antenna theory, and is often found in references called array gain for transmitting elements.

An array of sensors is more beneficial to the detection of underwater targets because of improvements in the signal-to-noise ratio as compared to a single sensor element. The directivity factor, K, is the ratio of the SNR of the array to the SNR of one omnidirectional element of the array.

$$K = \frac{\left(\frac{S}{N}\right)_{\text{directional}}}{\left(\frac{S}{N}\right)_{\text{omnidirectional}}} \qquad (10\text{-}19)$$

If the assumption is made that the signal developed by the array along the directional acoustic axis is equal to the signal that is distributed omnidirectionally ($S_{\text{directional}} = S_{\text{omnidirectional}}$), K then becomes:

$$K = \frac{N_{\text{omnidirectional}}}{N_{\text{directional}}} \qquad (10\text{-}20)$$

Omnidirectional noise ($N_{\text{omnidirectional}}$) is defined as $4\pi r^2 I_i$, and directional noise, is defined as

$$N_{\text{directional}} = \int_0^{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} I_i r^2 b(\theta,\phi)\cos\theta \, d\theta \, d\phi \qquad (10\text{-}21)$$

where the function $b(\theta,\phi)$ is the beam power distribution pattern of the array in polar coordinates. If the beam pattern has rotational symmetry and is nondirectional in the ϕ axis such that $b(\theta,\phi) = b(\theta)$, K simplifies to:

$$K = \frac{4\pi}{\int_{\frac{-\pi}{2}}^{\frac{\pi}{2}} b(\theta)\cos\theta \, d\theta} \qquad (10\text{-}22)$$

The function $b(\theta,\phi)$, for simple arrays, can be found and evaluated mathematically. Some common expressions for K are given in table 10-1. The directivity index (DI) is then simply the directivity factor ratio expressed in decibels.

$$DI = 10\log K \qquad (10\text{-}23)$$

TABLE 10-1 *Directivity index for various common arrays*

Type	DI = 10 log K
Continuous line of length L where $L > \lambda$	$K = \frac{2L}{\lambda}$
Piston of diameter D where $D > \lambda$	$K = \left(\frac{\pi D}{\lambda}\right)^2$
Line of n elements of equal spacing d	$K = \frac{n}{1 + \frac{2}{n}\sum_{i=1}^{n-1} \frac{(n-i)\sin\left(2i\pi d/\lambda\right)}{2i\pi d/\lambda}}$

Target Strength

Of special interest in the active sonar equation is the echo returned by an underwater target, called target strength (TS), and the fact that it usually is on the order of 15dB to 25dB. Echo returns may be of military value such as a submarine or mine, or the returns may be undesired echoes such as a school of fish.

The electromagnetic corollary is a radar cross section, and correspondingly shape, size, aspect, and construction material are crucial to the amount of TS an object has. The logarithmic value of the ratio of the intensity of sound returned by the target at a distance of 1 meter in some direction to the incident intensity from a distant source is called TS.[iv]

$$TS = 10 \log \frac{I_r}{I_i} \qquad (10\text{-}24)$$

where:

I_r = intensity of return at 1 meter

I_i = incident intensity

The variability of TS is a function of the target aspect presented to the incoming signal. A beam-aspect target presents a greater reflective area than one with a bow aspect; hence more energy is reflected from a beam-aspect target. Because TS is ten times the logarithm of the reflected intensity divided by the inbound intensity, this statement apparently says that more energy is reflected than is incident, a condition that is clearly not possible.

The key lies in the definition of terms:

I_r is the intensity of the reflected signal measured 1 meter from the target, assuming the target is a point source.

I_i is the intensity of the signal inbound from the transmitting ship to the target measured at a point on the target.

Intensity is actually power per unit area striking the target at some point, and thus the total sound power striking the target is I_i times an effective area. If one assumes that the major portion of this power is reflected not from the original effective area, which is almost the same as the profile area of the target, but instead from a point source, it necessarily follows that the reflected energy computed in this way must be greater because of the reduced area from which the energy emanates. Thus I_r is greater than I inbound if both are defined as indicated above.

In this case, there is no such wave as the one that is defined as originating from the point source. This construct is merely a convenient way of duplicating the actual measured value of I_r when the wave is 1,000 meters or more away from the point source en route back to the transmitting sonar.

Thus, if one were to measure I_r and I_i both at 1,000 meters from the target, then I_i would definitely be greater than I_r. Therefore, I_r would have been computed to suffer a greater attenuation in traveling 1,000 meters from the constructed point source than I_i would suffer in going 1,000 meters to the target.

The explanation for this is the rapid attenuation due to spreading from the point source as compared to that undergone by the inbound wave, which is very near a plane wave when it is within 1,000 meters of the target.

Noise Level

Noise level (NL) used in the sonar equations is the total contribution of the noise at the particular frequency of interest by both ambient ($NL_{ambient}$) and detecting platforms (NL_{self}). Chapter 9 contains a detailed analysis of these sources and how they are combined into one term.

Reverberation Level

Reverberation is the term applied to returns reflected by objects other than the desired target. Examples of such objects are entrained air bubbles and particles, fish, the ocean surface, and the ocean bottom. The echo created by the immediate surrounding volume of water is called reverberation, and this phenomenon limits the maximum source level. Therefore, the reverberation level (RL) increases with an increase in the active source level. The reverberation always comes back from the same direction in which the sonar is projecting.

Passive Sonar Equation

A passive sonar depends on receiving a signal that is radiated by a target. The target signal can be caused by operating machinery, propeller noise, hull flow noise, and so on, but the same fundamental signal-to-noise ratio requirement must be satisfied.

At the receiver, the passive equation begins as $S - N \geq DT$. If the target radiates an acoustic signal of SL (target source

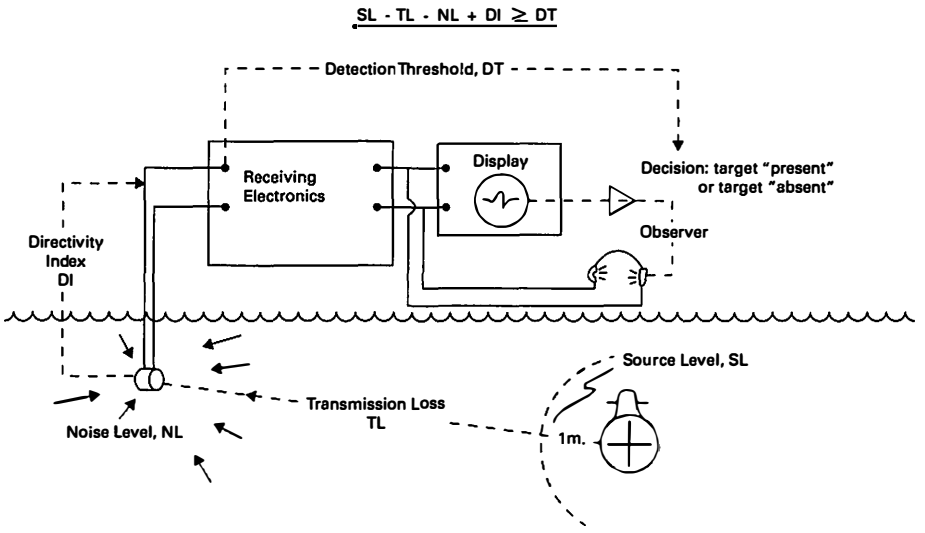


FIGURE 10-5 *Diagram of the passive sonar equation*

level), the sound intensity is diminished while en route to the receiver because of any one or more of the following: spreading, ray path bending, absorption, reflection, and scattering.

The decrease in intensity level due to this is transmission loss (TL) measured in decibels. Hence the intensity level of the signal arriving at the ship is:

$$S = SL - TL \qquad (10\text{-}25)$$

Noise, N, acts to mask the signal and is not wanted. Therefore, the receiver is composed of many elements that are sensitive primarily in the direction of the target so that the receiver can discriminate against noise coming from the other directions. This discrimination against noise is the directivity index, DI. The DI gives the reduction in noise level obtained by the directional properties of the transducer array. Therefore, in the basic equation, noise is now reduced and becomes:

$$N = NL - DI \qquad (10\text{-}26)$$

There are two things to note in this simple equation: the first is that the DI is always a positive quantity, so NL – DI is always less than or equal to NL; the second is that the parameter NL represents both the self-noise level (NL_{self}) and the ambient the noise level ($NL_{ambient}$), by its definition because it is the noise at the hydrophone location and can come from any, or all, sources.

The passive sonar equation can now be constructed in terms of signal and noise. When S and N are substituted from equations 10-25 and 10-26 into 10-1, the result is:

$$SL - TL - NL + DI \geq DT \qquad (10\text{-}27)$$

which is the simplest form of the passive sonar equation.

In words, equation 10-27 says that the source level of the target minus the loss due to propagation through the medium, minus the sum of all interfering noises, plus improvement by the spatial processing gain of the receiver, must be equal to or greater than the detection threshold for a target to be detected with the specified probability of detection.

However, the greater-than or equal-to condition is normally written as an equality. It is then understood to mean that if the left-hand side's algebraic sum is greater than DT, detection is possible with a greater probability than that specified by DT. If the sum is less than DT, detection probability decreases. Generally speaking, these two conditions imply that either detection is highly probable or it seldom occurs. As a further aid to understanding the passive sonar equation, figure 10-5 illustrates the sonar parameters and indicates where each term involved interacts to produce the desired results.

Active Sonar Equation

In an active sonar, acoustic energy is transmitted, and the received signal is the echo from the target. Two different, but related, equations are needed to describe the active sonar, one for an ambient noise-limited situation and the other for the reverberation-limited situation. As discussed previously, sonar performance is governed by the requirement that signal minus noise must be equal to or greater than the detection threshold.

The difference in the two active sonar equations that satisfy this requirement depends upon the characteristics of the noise that is actually present at the receiver when the signal is detected. The ambient noise may be described as either isotropic—that is, as much noise

power arrives from one direction as from any other—or as reverberation, in which noise returns primarily from the direction in which the sonar has transmitted.

Noise Limited

The development of the active sonar equation is similar to that for the passive equation. In other words, the formal sonar parameters will be fitted to the signal and noise terms of equation 10-1.

If a sonar transmits an acoustic pulse with an initial source level of SL in dB, the transmitted pulse will suffer a transmission loss in traveling to the target. The target will scatter acoustic energy, some of which will return to the sonar. The backscattered intensity is called target strength, TS, and as mentioned earlier is related to the scattering cross section of the target. The returning echo will again undergo a propagation loss, and thus the signal at the sonar will be:

$$S = SL - 2TL + TS \qquad (10\text{-}28)$$

As long as the source of the radiated energy and the receiver for the echo are located together, the transmission loss experienced is equal to twice the one-way transmission loss.

When the echo returns, under some conditions the reverberation background due to the initial transmission will have disappeared, and only ambient noise will be present. This noise will be identical to that described in the passive sonar equation, modified by the receiving directivity index. The fundamental relationship can then be expressed as

$$SL - 2TL + TS - NL + DI \geq DT \qquad (10\text{-}29)$$

which is the basic active sonar equation used when the sonar is operating in a noise-limited situation.

Reverberation Limited

If, on the other hand, the echo returns when the reverberation background has not decayed to a level below the ambient noise level, the background noise is given by RL. In this case, the parameter DI, defined in terms of an isotropic background, is inappropriate, inasmuch as reverberation is by no means isotropic.

For a reverberation background, the term NL – DI is replaced by an equivalent reverberation level observed at the transducer terminals, and the sonar equation takes the form:

$$SL - 2TL + TS - RL \geq DT \qquad (10\text{-}30)$$

which is known as the reverberation-limited active sonar equation.

Detailed quantification for the new term, RL, is difficult at best, because it is a time-varying function resulting from the inhomogeneties in the medium.

One thing to note is that in the normal conversion from the basic equation to the active equations, the inequality again becomes an equality. As discussed under the passive sonar equation section, it is understood that when the terms on the left-hand side exceed the detection threshold by a significant amount, detection is highly probable, and when the terms are significantly less than the detection threshold, detection seldom occurs. Figure 10-6 pictorially depicts the active sonar equations.

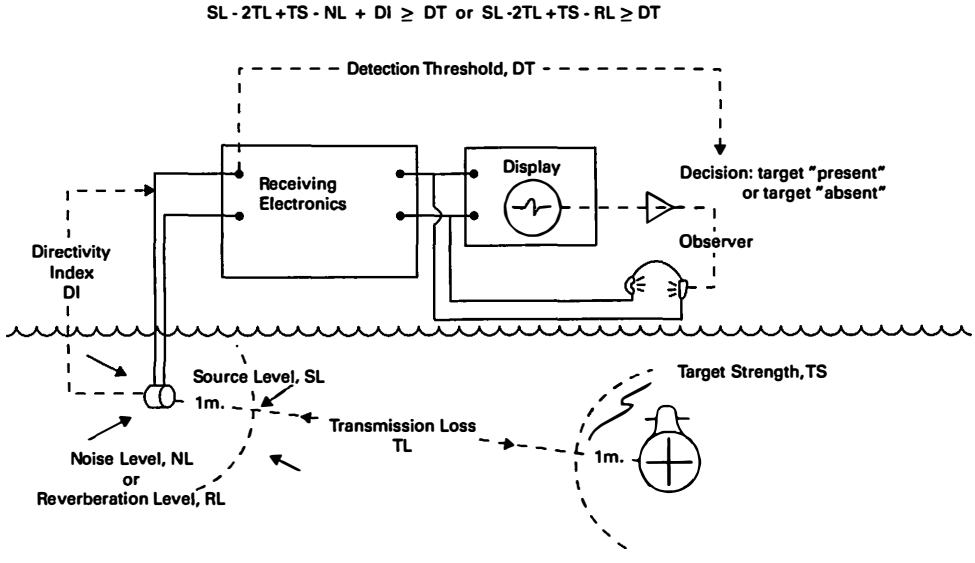


FIGURE 10-6 *Active equation system*

Figure of Merit

The possible detection range of particular equipment should be known so that a tactician will then have a measure of the sonar's capability and an understanding of what the sonar can do in a given tactical situation. Unfortunately, with no change in basic sonar configuration, its detection capability measured in terms of range can increase or decrease several-fold simply because the ocean itself has changed.

To state this another way, sonar equipment can only be designed to detect the arrival of a certain intensity of sound. The range from which that sound intensity arrives is highly dependent on how much energy was lost en route, and therefore, detection range alone is a poor measure of sonar capability.

A better measure is the ability of the sonar to detect a certain level of sound energy intensity just outboard of its receiver. The key to this better measure of performance is separating the sonar from the ocean in which it must operate. Only then can sonar capability be discussed in terms of the unchanging sonar hardware as distinguished from the ever-changing ocean.

This better measure for sonar capability is called figure of merit (FOM), and it equals the maximum allowable one-way transmission/propagation loss in passive sonars, or the maximum allowable two-way transmission/propagation loss in active sonars for a detection probability of 50%. Therefore, solving equations 10-27, 10-29, and 10-30, for transmission loss, we get:

$$\text{passive FOM} = \text{SL} - \text{NL} + \text{DI} - \text{DT} \qquad (10\text{-}31)$$

$$\text{active FOM (noise limited)} = \text{SL} + \text{TS} - \text{NL} + \text{DI} - \text{DT} \qquad (10\text{-}32)$$

$$\text{active FOM (reverberation limited)} = \text{SL} + \text{TS} - \text{RL} - \text{DT} \qquad (10\text{-}33)$$

This combination of terms is probably the most-used performance parameter for sonars, and it is important to understand just what it means. The FOM of a sonar system is the maximum loss that can be tolerated and still provide the necessary detection probability as specified by the DT. The FOM is improved by raising the source level, which can be accomplished by increasing the transmitted power in the active case or finding a noisier target in the passive case, decreasing the ambient noise level, increasing the absolute value of the receiving directivity index, and decreasing the detection threshold.

The importance of the figure of merit is that with no knowledge of the intervening propagation path between a ship and a target, a quantitative comparison of two different sonars or conditions can be made. The FOM may be utilized for comparing the relative performance of two passive sonars, provided the calculations of the comparative figures of merit are made for the same frequency. The differences in FOMs represents additional allowable propagation loss that can be sustained by the sonar having the higher FOM and still making a detection; thus a longer detection range results.

Propagation Loss Curves

Naturally, to the tactician, detection ranges are of prime importance, and although the FOM can be interpreted in terms of range, this can be done only if the propagation losses involved are known. The measure of propagation loss (total transmission loss) in many parts of the ocean has shown that not only does sound propagation change with frequency and location, but it changes with season as well.

Thus, the sound velocity profile will determine the propagation path(s) available, which, along with frequency, source, and receiver geometry as well as location, will determine transmission loss versus range. Propagation loss curves can be made from ray tracing computer algorithms or actual measurements and smoothed data. Therefore, in order to convert the FOM to range, one must have a propagation loss curve for the frequency of the sonar(s) concerned and for the area and season of the year for which range prediction is desired.

Figure 10-7 is a typical propagation loss curve for a frequency of 2 KHz. Note that the propagation losses for sound paths previously described are plotted. Multiple bottom bounce losses are shown and multiple convergence zones are estimated.

EXAMPLE 10-1: Assume two different passive sonars with FOM values at 2 KHz of 80dB and 90dB. Based on figure 10-7, the range in Kyds at which 50% probability of detection given the sound transmission path is predicted for each sonar in table 10-2. As shown, the higher the FOM, the longer the detection range for a given path.

TABLE 10-2 *Predicted ranges at different FOMs based on the propagation loss curve in figure 10-7*

Sonar (FOM) dB	Direct Kyds	Bottom Bounce Kyds	Convergence Zone Kyds
80	18	none	none
90	23	36–47	81–96

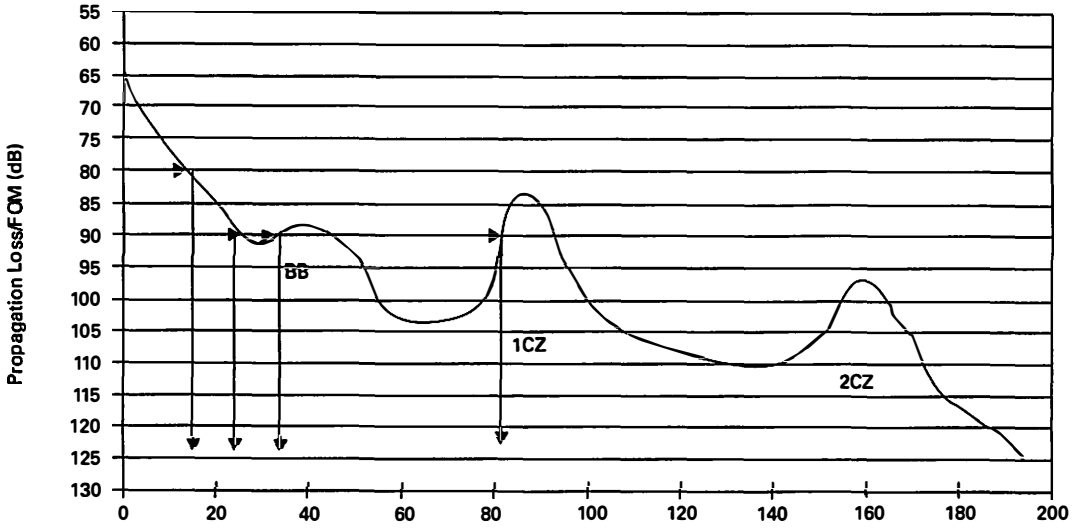


FIGURE 10-7 *Propagation loss curve*

Note that the sonar with the higher figure of merit permits the use of different sound paths.

In summary, sonar performance is key to undersea warfare success, and the figure of merit is the key to sonar performance. With knowledge of the sonar's FOM, a commanding officer can ensure that equipment is peaked, and also predict detection ranges against possible enemy targets. The war planner can do likewise for either a real or a hypothetical enemy.

Because the changing ocean causes dramatic changes in propagation loss versus range, stating a sonar's capability in terms of range tells only half the story, and may even be misleading. Using figure of merit, however, means that the sonar with the higher FOM will always be the better sonar when compared to another sonar in the same mode.

EXAMPLE 10-2: Figure of Merit. Your sonar is capable of either passive or active operation. You are operating in the shipping lanes with a sea state of 2. Water depth is 200 fathoms. Using the following information, you must decide which mode to use. Intelligence information indicates that the threat will be a Zebra-class submarine (all dB are reference 1 μPa).

Target parameters:

Radiated noise source level	100dB
Radiated noise frequency	500 Hz
Target strength	15dB
Target detection range	10,000 m

Sonar parameters:

	Active	Passive
Source level	180dB	
Frequency	1.5 KHz	
Self-noise at 15 kts	50dB	50dB
Directivity index	10dB	8dB
Detection threshold	2dB	3dB

SOLUTION: In order to determine which sonar to use, it is necessary to calculate the FOM and total transmission loss for each mode.

First, calculate the total transmission loss for each mode. Since the desired target detection range of 10 kilometers is much greater than the water depth of 200 fathoms, we will use equation 10-18 for cylindrical spreading:

$$TL = 10\log r + 30 + \alpha r + A$$

The quantity A is assumed to be zero since no information is available. The absorption coefficient (α) is calculated by substituting signal frequencies for each mode into equation 10-15:

$$\alpha = \frac{0.036 f^2}{f^2 + 3{,}600} + 3.2 \times 10^{-7} f^2$$

Active Mode

$$\alpha = \frac{0.036(1.5)^2}{(1.5)^2 + 3{,}600} + 3.2 \times 10^{-7}(1.5)^2 = 2.32 \times 10^{-5}$$

$$TL = 10\log(10{,}000) + 30 + 2.32 \times 10^{-5}(10{,}000)$$
$$= 70 + 0.232$$
$$TL = 70.232\text{dB} \quad \text{(one way)}$$
$$2 \times TL = 140.46\text{dB (Total TL)}$$

Note: TL depends only upon the detection range and the frequency of the signal.

Next calculate the FOM using equation 10-32 for the active mode and equation 10-31 for the passive case. *Note:* the only values that must be determined are the noise levels using the Wenz curves and nomograms for the different frequencies.

Passive Mode

$$\alpha = \frac{0.036(0.5)^2}{(0.5)^2 + 3600} + 3.2 \times 10^{-7}(0.5)^2 = 2.6 \times 10^{-6}$$

$$TL = 10\log(10{,}000) + 30 + 2.6 \times 10^{-6}(10{,}000)$$
$$= 70 + 0.026$$
$$TL = 70.026\text{dB(Total TL)}$$

Active Mode

f = 1,500 Hz

NL(shipping) = negligible
NL(sea state) = 58dB
NL(self-noise) = 50dB

Combining the signals for each mode:

$$NL = 10\log\left(10^{58/10} + 10^{50/10}\right)$$
$$= 58.64\text{dB}$$

Determining the FOM is now a simple matter of substituting the calculated and given values into the appropriate equations:

Passive Mode

f = 500 Hz

NL(shipping) = 57dB
NL(sea state) = 61dB
NL(self-noise) = 50dB

$$NL = 10\log(10^{57/10} + 10^{61/10} + 10^{50/10})$$
$$= 62.70\text{dB}$$

Active Mode

FOM = SL + TS − NL + DI −DT
FOM = 180 + 15 − 58.64 +10 − (−2)
FOM = 150.36dB

Compare the FOM of each mode with the total TL for each mode to determine which mode is optimum for this target. The FOM for the active case is greater than the total TL. Therefore, the active mode will give your ship greater than 50% probability of detection. The FOM for the passive case is less than the total TL. Therefore, the passive mode will give a less than 50% probability of detection, which means that the active mode should be used versus the passive.

Passive Mode

FOM = SL − NL + DI − DT
FOM = 100 − 62.70 + 8 − 3
FOM = 42.30dB

[i] Hall, 194

[ii] Urick, 153

[iii] FAS, sec. 8.7.3.1

[iv] Urick, 291

11 Undersea Warfare Systems

Introduction

Various sensor systems and platforms are involved in undersea warfare. Each sensor system—whether located on surface, subsurface, airborne, or space platforms—has its strengths and weaknesses. Because of the complexity of the medium involved, undersea warfare uses all means available to detect underwater ships and other objects. Therefore, this chapter will present the various detection systems available and the basic fundamentals of each of these systems. Through such system knowledge, personnel engaged in undersea warfare can maximize their chances of success.

Transducers

A device for converting one form of energy to another is a transducer. In sonar, electrical energy is converted to acoustic energy in the form of oscillations in the water molecules through which the sound travels. These devices transform energy between electronic and acoustic form in either direction. They also preserve the structure of the waveforms during their transformation, and can be approximated to be reversible and linear; see figure 11-1.[i]



FIGURE 11-1 *Transducer functions*

A transducer that converts electrical signals to sound is called a projector, whereas a hydrophone only converts sound into an electrical signal. In some sonar applications, the same transducers are used for both the source and the receiver functions whereas others use separate projectors and hydrophones.

Three physical phenomena that exhibit the ability to change electrical energy to mechanical energy and mechanical energy to electrical energy are employed in sonar transducers. These are the piezoelectric, electrostrictive, and magnetostrictive effects.

Materials exhibiting piezoelectric and electrostrictive properties are generally of crystalline or ceramic nature. They change dimensions when subjected to an electric field and develop a voltage potential between two opposite faces when mechanically stressed.

Materials exhibiting the magnetostrictive effect change dimensions when subjected to a magnetic field and change magnetic permeability when subjected to mechanical forces. This change in permeability alters the intensity of any magnetic field in the presence of the material.

Crystal Transducers

The piezoelectric effect is the electric polarization produced by mechanical strain in certain classes of crystals. When placed under pressure, crystalline substances such as quartz, ammonium dihydrogen phosphate, and Rochelle salt acquire both a charge and a stress when a voltage is placed across them.[ii] The polarization and the electric potential induced by it are proportional to the strain and changes sign with it. In the converse piezoelectric effect, an electric potential across the crystal face produces a mechanical deformation proportional to the induced electric polarization and of the same sign. The magnitude of this effect varies for crystals of different material, and for the different axes of the crystal; see figure 11-2.

A crystal transducer is constructed by cementing or mounting a stack of quarter-wavelength crystals to a heavy steel backing resonator plate that is also one quarter-wavelength thick; see figure 11-3. The assembly is mechanically resonant at the operating frequency. When used for listening, the crystals produce the maximum electric signal as long as the frequency of the received sound is the same as the resonant frequency of the transducer.

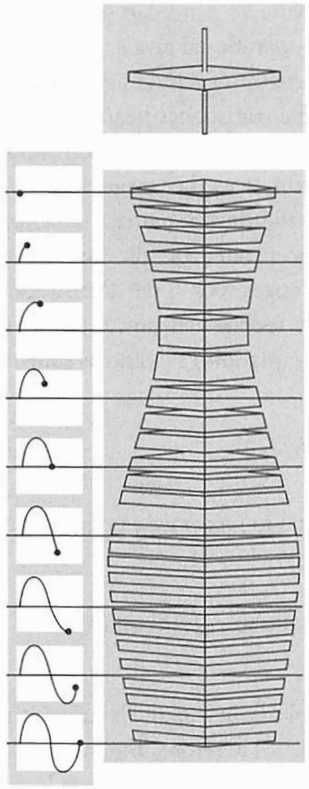


FIGURE 11-2
Piezoelectric effect—AC current is applied to a crystal, causing vibration.

Ceramic Transducers

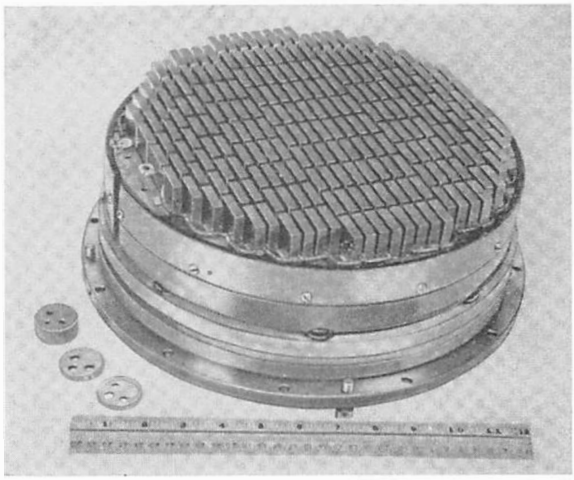
Electrostrictive transducers are made from polycrystalline ceramic dielectric material such as barium titanate and lead ziconate titanate. As the ceramic material is being cooled, an intense constant electric field is applied to the material to properly polarize the transducer.[iii] After polarization and after an electric field is applied to the dielectric, the dielectric is deformed.

As an AC current is applied to a transducer, it vibrates at the same frequency as the current. This phenomenon of change in dimension is called electrostriction and is independent of the direction of the electric field and proportional to the square of the field intensity.

Alternately, if a mechanical force is applied to the dielectric, a voltage is produced with the same frequency as the applied force. Ceramic transducers have an advantage over crystals in that the ceramic can be molded to any desired shape. This property is particularly desirable for making cylindrical scanning or omnidirectional transducers used in active sonar systems.

Magnetostrictive Transducers

A magnetic field will cause a number of materials to change dimensions in the direction of the applied field. Depending on the material and the strength of the magnetic field, some special materials will expand and some will contract, but independently of the direction of the applied field. This is the magnetostrictive effect.



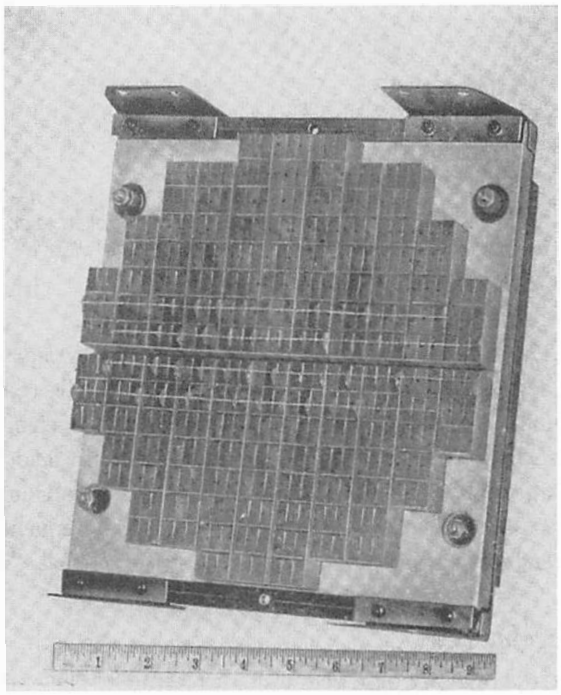


FIGURE 11-3 *Crystal stack for a transducer*[iv]

In one type of transducer, nickel rods are placed in a continually varying magnetic field, which causes the rods to vibrate. The frequency of the vibration is dependent on the frequency of the AC voltage applied.

The magnetostrictive effect also works in reverse. When the rod is caused to vibrate by some outside source such as impinging sound waves, a varying magnetic field and AC voltage are produced. To increase the area of the active face in contact with the water in this design, a series of tubes or rods is attached to a diaphragm, and the diaphragm becomes the active face of the transducer.

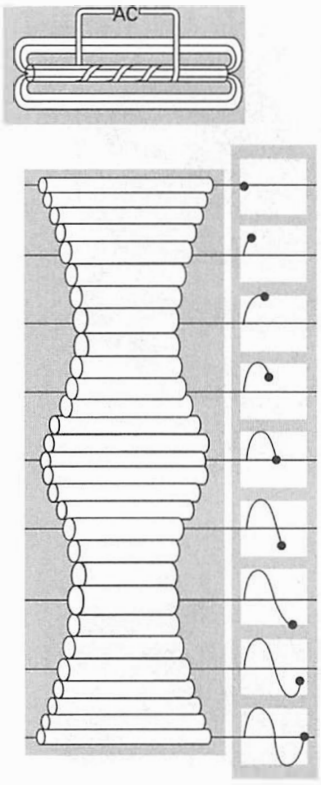


FIGURE 11-4
Magnetostrictive effect—AC voltage applied to coil, causing tube of vibrate

The dimensions of the tubes and the diaphragm plate are designed to make the transducer mechanically resonant at the operating frequency for maximum efficiency. In such a transducer, the sinusoidal magnetic field is obtained by winding coils around the nickel rods and impressing a current at the desired sonar signal frequency through the coils; see figure 11-4.

Hydrophones

Hydrophones are designed only for reception. They may be either magnetostrictive or electrostrictive and use the same principles as two-way transducers. However, since they do not handle high-transmission power, with its resulting increase in temperature, their construction is generally much lighter.

Transducer Directivity

The directivity or beam pattern of a transducer is a function of the mechanical arrangement of the transducer elements and the associated electrical and electronic circuitry. As with radio and radar antennas, where the directivity pattern is caused by interference between electromagnetic radiation from various parts of the antenna, in transducers the patterns are caused by interference between sound radiated from various points on the transducer surface. Control of transducer beam patterns is needed so that the directivity can be matched to operational needs.

Transducers are also designed to exclude noises from directions other than along the transducer beam by suppressing back- and sidelobes, thus reducing background noises and permitting detection of weaker target signals.

Additionally, control of the beam pattern enables transmitted power to be concentrated on the target for stronger echoes and longer ranges. For a flat transducer with a circular face, the angle θ for the half-power point of the main beam is predicted by equation 11-1, and as shown in figure 11-5:

$$\sin(\theta) = 1.6\lambda/2\pi r \qquad (11\text{-}1)$$

where:

θ = angle from the normal to half-power point (deg)
λ = sonic wavelength
r = transducer radius
1.6 = half-power point of the Bessel function[v] $[2J_1(x)/(x)]$

For satisfactory operational resolution, the beamwidth must not be greater than 10°. Therefore, the ratio r/λ must be at least 3, and from this it can be seen that low-frequency operations require large-size transducers.

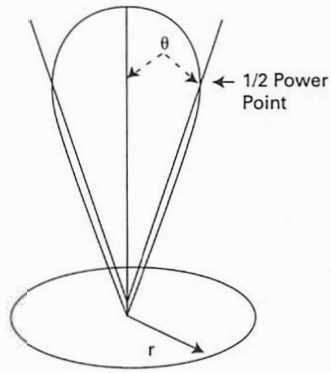


FIGURE 11-5 *Flat plate half-power point beamwidth*

Directivity patterns can be modified for the reduction of sidelobes by shading a transducer. A transducer is shaded in design if the power projected per unit area is varied over the surface. Shading can be accomplished:

1. Physically, by arranging the spacing, size, or number of elements.
2. Electrically, by the number of turns on magnetostrictive elements, or by phasing and delay networks.
3. Mechanically, by the physical shape of the active surface of the transducer.

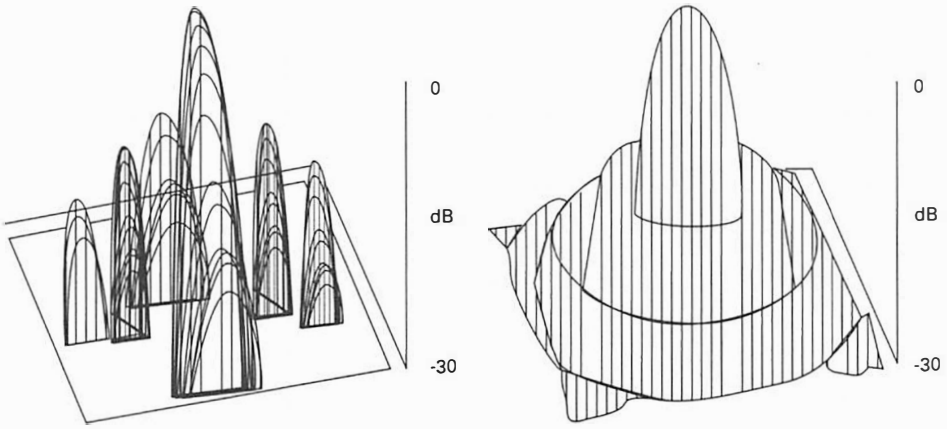


FIGURE 11-6 *Three-dimensional beam patterns for rectangular and circular projectors*[vi]

Direction Determination

To be useful, a sound signal must not only be produced and detected, but the direction of the signal must also be determined. For this discussion, assume that a sound wave is from a distant source such that the wave front is parallel. Note also that the following principles hold for the generation of an acoustic signal.

When a sound wave from a distant source strikes the transducer front such that all elements of the transducer vibrate in phase, this means that the source is on the boresight of the transducer face. However, when the array is not pointed directly at the sound wave (but is inclined to the wave front), the signals produced by the elements are out-of-phase, resulting in a signal that is something less than maximum; see figure 11-7. Because the sound wave hits each element at a different time, the signals produced by the elements begin at different times and, thus, are out-of-phase. By building in delay circuits that alter the phase of each element electronically, the direction of the signal is determined by adjusting the phase of each element until a maximum signal is achieved.

Transducer Power

A transducer transforms electrical energy into mechanical energy and then into acoustic energy by generating an alternating pressure signal that is superimposed on the static ambient pressure of the water through which the sound is propagating. The relationship between acoustic power output, P, of the transducer and the sound pressure is:

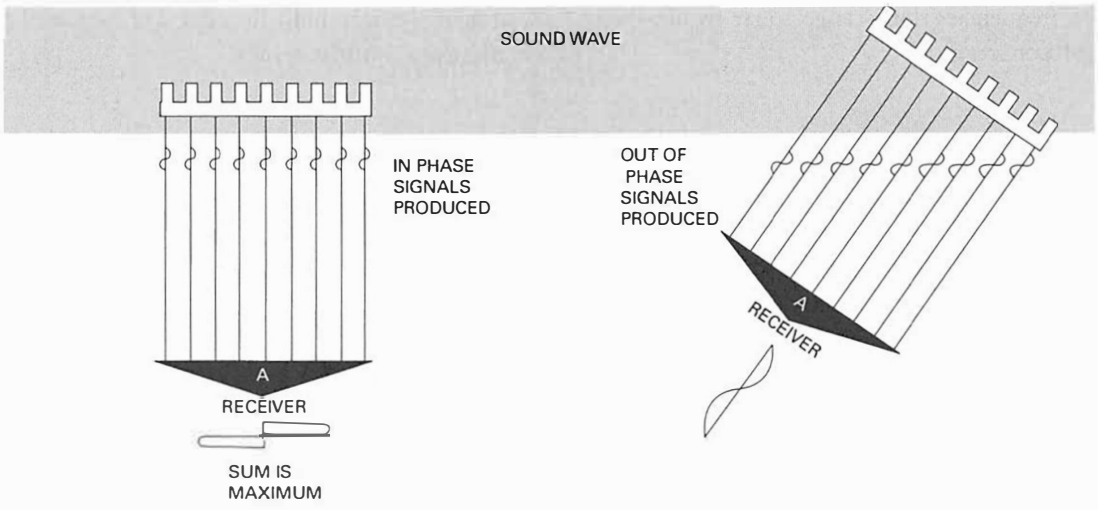


FIGURE 11-7 *Directionality of a transducer*

$$P = P_e^2(A/\rho c) \qquad (11\text{-}2)$$

where:

P = acoustic power output (w)
P_e = rms sound pressure at 1 meter (Pa)
A = area of transducer through which the acoustic energy flows (m^2)
ρ = density of water (kg/m^3)
c = sound speed of water (m/s)

The pressure of the sound wave generated at the surface of the transducer is limited by the hydrostatic pressure at the transducer, which in turn is limited by the cavitation pressure, the pressure below which water vaporizes.

If the transducers are driven with too much power, they can cavitate by dropping the pressure so low that the water boils. This is called quenching, and it can destroy the transducer since the normal backpressure is removed when bubbles form on its surface. Since the normal restoring force is gone, the surface of the transducer can travel too far (overrange) and damage itself. This factor limits maximum transmitted power to about 0.3 watt per square centimeter of the transducer area in the region near the surface of the sea.

The quenching power limit increases with depth due to the increased ambient pressure as the square of the ambient pressure. Thus, at the depth where the ambient pressure is doubled, the maximum transmitted power becomes about 1.2 watts per square centimeter of transducer area.

Increasing transmitted power for a given transducer configuration thus normally means increasing transducer array size. Transducer array size is also directly related to signal frequency, and for a reasonable beamwidth, the transducer array face must be six to eight wavelengths wide. Since maximum range is obtained with lower frequencies, long-range sonar requires very large transducer arrays.

This size versus power and frequency situation presents severe physical and tactical limitations. Large transducer arrays add considerably to drag effects even with streamlined domes. They also present problems on how and where to mount them, and require very large power sources that further require large ships to accommodate their associated electrical and electronic equipment. The total amount of power transmitted by a transducer array of fixed size and frequency is limited; therefore, concentrating the power in a beam by control of directivity patterns becomes the more practical method of increasing power.

Active Systems

Active sonar systems most closely resemble the operation of basic radar. Active sonar emits a short burst or pulse of sound energy called a ping that is transmitted through the transducer into the water. Range is measured by accurately determining the time interval between the transmitted burst and the instant its reflection returns from the target or some discontinuity in the water. Bearing is determined using beams and phase differences in the return signal as it strikes the transducer face. A functional diagram of an active sonar system is shown in figure 11-8.

Transducer Array

The transducer array is simply multiple transducers arranged in a fashion to improve directivity and thus the FOM. One transducer alone has little or no directionality; however, when several transducers are spaced apart at known distances, special signal-processing techniques are applied to form narrow beams. For surface ships, beams are formed to look out and down instead of straight ahead to limit the effects of increased noise and reflections near the surface.

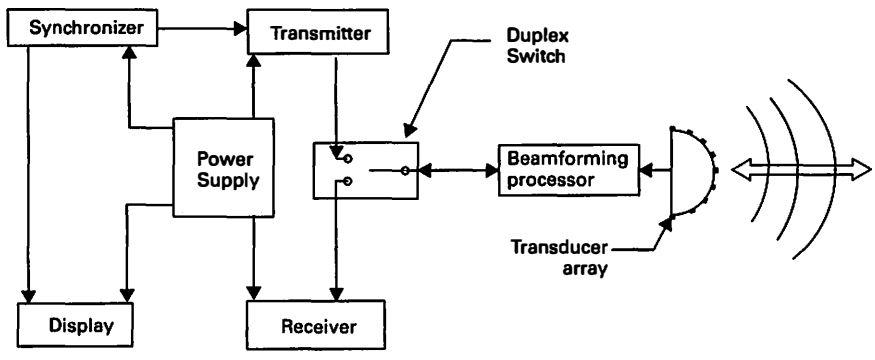


FIGURE 11-8 *Active sonar system*

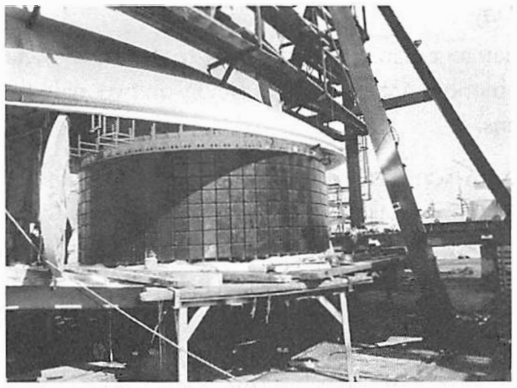


FIGURE 11-9 *Cylindrical array of transducer elements with rubber sonar dome removed from bow; note the baffles directly aft of the array.*

The normal configuration of hull-mounted arrays is usually cylindrical to give complete coverage out and down except in the baffles; see figure 11-9 and 11-10. Baffles are the region directly behind the array that is designed to protect the occupants of the ship from the loud sound source of the sonar and is made with sound-absorption material to prevent reflection back into the system. Because of this, target detection cannot occur in the baffle region.

Sonar equipment mounted to the hulls of ships presents various problems. Self-generated noise due to water action around the sonar dome requires extreme consideration in the design of the dome to provide adequate streamlining. Streamlining is also required to minimize drag. The drag of the sonar dome, particularly for large long-range sonar, can be considerable, so in some applications, retractable domes are used.

Shielding is required to block off own-ship machinery noise, and flow noise. and vibration mounts may be needed to isolate the transducer from shipboard vibrations. The problem of interference from own-ship hull noise with sonar operations has led to placing the transducer in a streamlined dome at the bow. The disadvantage of this application is that the bow dome can be easily damaged and thus requires careful maneuvering during docking and high-sea states; but its acoustic advantages have resulted in this design being the most common in surface warships and submarines.

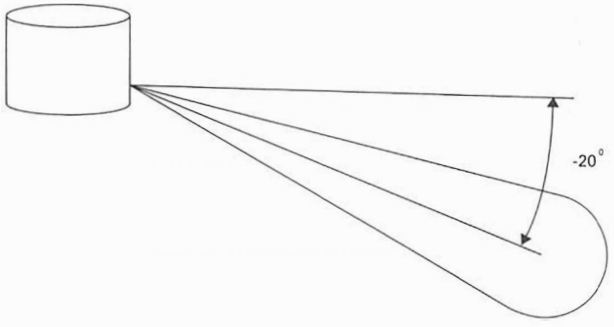


FIGURE 11-10 *Vertical beam of typical cylindrical transducer array (declination angle)*

Transmitter

The transmitter generates the outgoing pulse. It determines pulse width, PRF, modulation (optional), and carrier frequency. The pulse formed by the transmitter goes through the transmit/receive switching section (duplexer) and is applied to all elements of the transducer. Therefore, the transducer transmits a burst of sound of the same duration and frequency as the exciting pulse. The operator can control the output power or source level to limit quenching and to help avoid counterdetection.

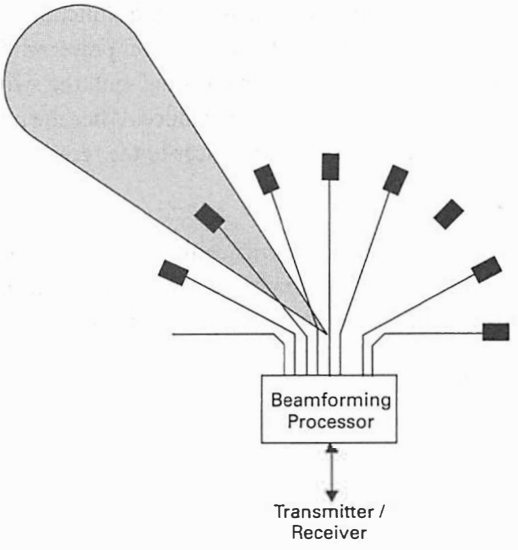


FIGURE 11-11 *Active beamforming*

Beamforming Processor

The input/output of each transducer is put through a beamforming processor, which applies time delays or phase shifts to each of the signals; see figure 11-11. This creates a narrow beam in a particular direction, much the same as the phased-array radar. In addition, when noise exists in a particular direction, a null can be placed in the beam pattern to cancel out the noise, thus increasing the signal-to-noise ratio. This cancellation process is called adaptive beamforming and is accomplished by using appropriate digital techniques that automatically cancel out the interfering noise.[vii]

The beamformer processor determines the beamwidth of the systems and, consequently, the bearing accuracy. A dual, beam-tracking system, much akin to the radar

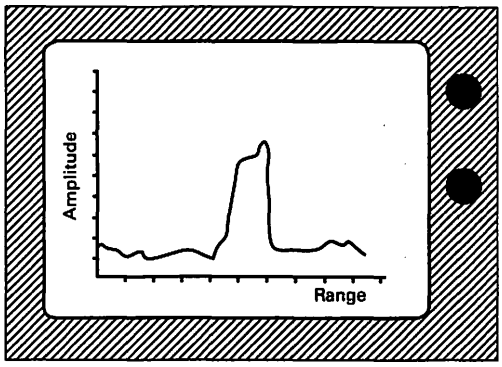


FIGURE 11-12 *A-scan display*

application, can improve bearing accuracy but at the expense of the SNR.

Duplexer

The duplexer is a switch that has the main function of isolating the receiver from the full transmitter power while the pulse is being sent. As the pulse is being sent, the switch is aligned to the transmitter and transducer. After the pulse is sent, the switch aligns the transducer to the receiver.

Synchronizer

As in a radar system, the synchronizer is the timer of the sonar. This device coordinates all transmit, receive, and display functions that require time as a reference.

Receiver

The receiver collects signals from the transducers for detection. Any object in the sound beam will reflect back to the transducer a portion of the sound energy. The transducer transforms this reflected sound energy into electrical energy. Again, the beamformer processor plays a large role in the signal supplied to the receiver.

The receiver amplifies the signal and compares the signal level to the noise level. To determine if the signal will be displayed in a particular beam, the signal level must exceed the predetermined detection threshold (DT) or threshold SNR. If the threshold is set too low, there will be many false alarms. If it is too high, some detection capability will be lost. The receiver may also demodulate the return if frequency modulation is used on transmission.

In addition, sonar systems often use pulse compression techniques to improve range resolution.[viii] To obtain an audible signal, the received signals are amplified and combined with the output of a local oscillator, akin to the superheterodyne receiver discussed in chapter 6, to give a beat frequency in the audible range of frequency. The audible output is then fed into a loudspeaker or earphones.

Display

Similar to radar, active sonar systems display detection information to the operator visually on two main types of screens.

1. A-scan-type displays plot range or time on the horizontal axis and signal amplitude on the vertical axis. A target in the beam of the pulse appears as a sharply raised section; see figure 11-12.
2. A plan position indicator (PPI) display shows a top-down view plotting range outward from the center and bearing (true or relative) clockwise; see figure 11-13. When sound is transmitted, energy can radiate in all directions from the transducer with equal intensity. Upon reception, scanning circuitry causes the transducer to be sensitive only to signals returning from a narrow sector.

   This sector is electronically sampled at high data rates, and the range and bearing information is presented on a cathode-ray tube. The use of this method makes it possible to obtain range and bearing on every target for every transmission; also, the intensity of the return on the cathode-ray tube is an indication of signal strength—the brighter the return, the stronger the signal.

Doppler

In the process of submarine detection, it is desirable to obtain as much information about the target as possible. Such information includes target motion, and it is here that sonar operators use the principles of the Doppler effect.

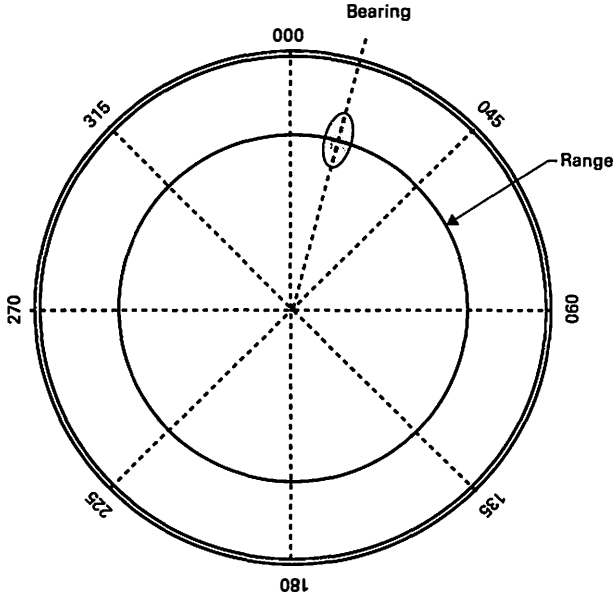


FIGURE 11-13 *PPI display*

The initial transmission is increased or decreased in frequency by the effect of relative motion. It strikes a target at this changed frequency and echoes at an altered frequency. As the sound returns, it is again affected by the same factors that acted upon the original sound, and consequently is changed in frequency again. Thus, it is received at the source with a frequency change double what normally would result from the given rate of relative motion.

Active sonar operators do not hear the outgoing sound transmission because the receiver is blocked during ping time. For this reason, a comparison of the outgoing transmission and the return echo cannot be made. However, a comparison can be made of the frequency of the echo from the target with that of the reverberations heard immediately before the echo is heard.

This comparison is extremely valuable because the difference between the pitch of the reverberations and the pitch of the target echo depends solely on the target's absolute motion in the path of the sound beam. Any change in transmission frequency that may result from source movement has its effect upon reverberations as well as upon the target echo; therefore the target can be detected.

An increasing pitch indicates that the target is heading toward the active source; conversely, a decreasing pitch indicates that the target is heading away. Both of these indications are independent of source motion and of the direction of range change. For example, a submarine may be closing in range due to the superior speed of the searching ship, but if the submarine is heading away from the ship, its echo will yield a decreased frequency because the target echo is being compared with the reverberations.

Actually, the frequency of both echoes and reverberations will be higher than the frequency of the original active transmission because of the closing range. But the increase in reverberation frequency will be greater than that of the target echo because the range rate between the source and the water particles causing reverberation is greater than that between the source and the target. The net effect therefore is a decrease in frequency that can be detected.

Doppler is valuable for more than the assumptions that can be made concerning a target's course and speed because a difference in frequency between two tones can aid in separating the target from the background noise. Even a very weak return can be distinguished if its pitch differs markedly from pitches of surrounding echoes. Thus, target Doppler is a great aid in detecting echoes against a background of reverberation.

Figure 11-14 shows the Doppler effect for several different situations. The curved lines indicate reverberations

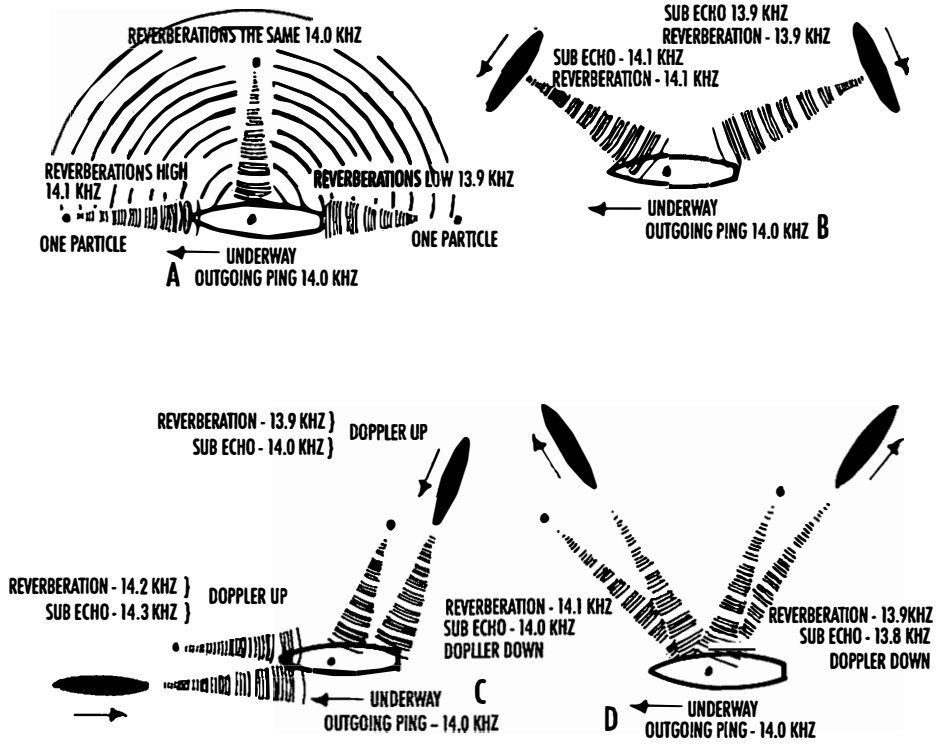


FIGURE 11-14 *Reverberation, transmitted pulse, and echoes*

and echoes returned from the ship (source). The encircled dots represent particles in the water that cause reverberations. In all views of the illustration, the transmitted pulse is 14.0 KHz.

In view A, reverberation alone is illustrated. Note the change in frequency caused by the motion of the ship.

In view B, the submarines are moving at right angles to the transmitted pulses; hence they cause no Doppler effect. Notice that the frequency of the echoes from the submarine is the same as the frequency of the reverberations.

View C illustrates two circumstances that produce up Doppler. Up Doppler is a sonar expression indicating that the target echo frequency is higher than that of the reverberations and that the target is in motion and heading toward the source.

View D is an example of down Doppler situations, when the target is in motion and is heading away from the source.

Doppler varies in its amount of frequency difference from the surrounding reverberation echoes in a manner dependent on the component of target speed in the line of sound. The value of Doppler thus depends on the speed of the target (and its aspect angle) by an amount equal to[ix]:

$$f_D = \frac{2V_{LOS}}{\lambda} \qquad (11\text{-}3)$$

where:

f_D = Doppler shift (Hz)
V_{LOS} = relative velocity in the line of sound (m/s)
λ = wavelength of the sound wave (m)

Doppler has another important application in the use of sonar, in addition to its estimating the factors needed for an accurate solution of the fire control problem. Many echoes are returned from floating debris, kelp, and other inhomogeneous particles suspended in the medium or at rest on the bottom. These echoes show no Doppler effect; they are motionless relative to the medium. Therefore, Doppler is a significant aid in the classification of an underwater contact as being or not being a submarine.

In this usage, however, a handicap arises. Stationary objects located in a current yield Doppler indications. A reef or pinnacle located in a current moving at 3 knots in a westerly direction, for example, appears to be underway at 3 knots in an easterly direction. This effect occurs because the target is not motionless with respect to the medium. Because the sound wave cannot differentiate between relative motion resulting from a moving target in a stationary medium and a stationary target in a moving medium, an erroneous indication of target motion is yielded. Conversely, a submarine drifting along with a current may yield no Doppler even though it is moving.

Passive Systems

The propellers, machinery, and various activities aboard ships and submarines produce sounds that are transmitted to the water, either directly from turbulence or through the ship's hull in contact with the water. These sounds are detected and identified at considerable distances with suitable equipment. The search for, detection of, and identification of noises produced by ships and submarines is termed passive listening.

Dedicated hydrophones, or the receive portion of an active sonar system, can be used for listening. Target bearings are determined in the same manner as for targets detected by the return of an active ping; however in a passive system, the target generates all the signals. Figure 11-15 shows a typical functional diagram of a passive sonar system.

It is important to note that active sonar equipment is needlessly complex when installing solely for listening purposes. Moreover, the frequency range over which

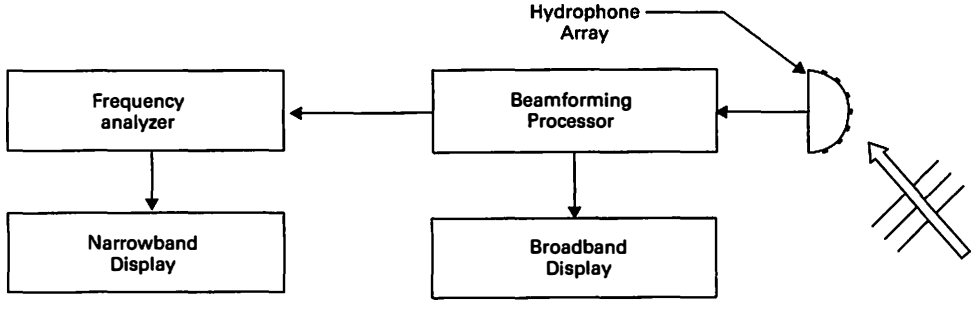


FIGURE 11-15 *Passive sonar system*

active sonar equipment has a good response is not optimum for general listening because it is designed to listen primarily at the transmitter frequency.

Hydrophone Array

Hydrophones, which are used exclusively for listening, are much lighter than transducers because they do not have to handle high power. They are generally designed with a flat signal response over a much wider frequency range than transducers. This frequency range is generally below that of transducers, partly because sound is attenuated less at lower frequencies and partly because the high-intensity sounds produced by machinery and propellers are in the lower frequency ranges. Ships and submarines can be detected at long ranges using this type of equipment.

Hydrophones are arranged into an array in order to improve the beamwidth and thus the directivity index. Common configurations are cylindrical, spherical, and conformal.

The cylindrical array is typically enclosed in the bow and consists of a series of hydrophones arranged vertically in a cylinder operating at a fixed vertical angle, usually downward. The spherical array, common on submarines, has hydrophones arranged on a sphere and has a much wider vertical field of view; see figure 11-16. Conformal arrays consist of groups of hydrophones mounted along the length of the ship or submarine.

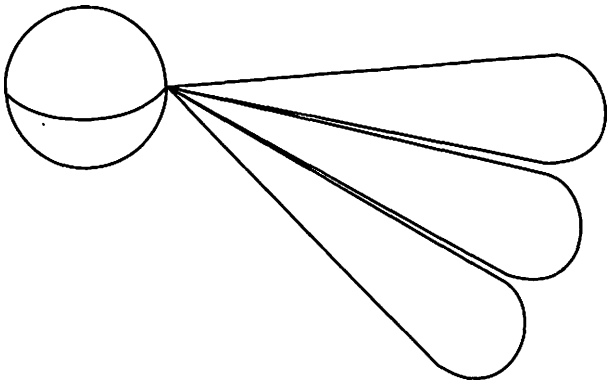


FIGURE 11-16 *Spherical array showing multiple vertical beams at various declination angles*

All of these arrays when mounted on a submarine must be capable of listening upward to some extent because a submarine may be below what the array is tracking. Using a beamforming processor, the field of view of the array is divided into individual beams in the vertical and azimuth directions.

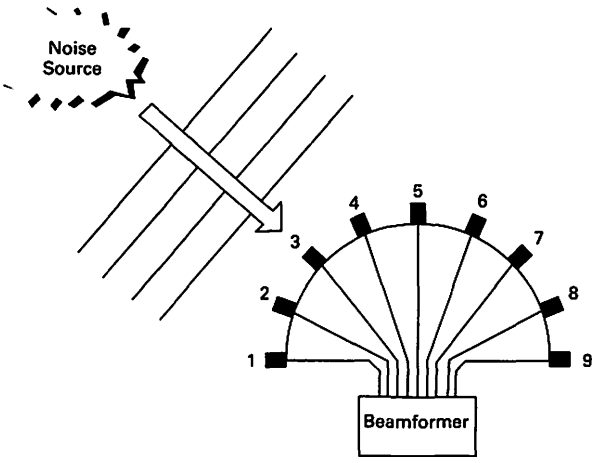


FIGURE 11-17 *Passive hydrophone array*

Beamforming Processor

Hydrophone arrays determine the bearing of the signal and enhance the detection of the received signals by increasing the signal-to-noise ratio. Both functions are performed by the beamforming processor, which applies time delays and summing of the signals from the hydrophone array.[x]

Unlike active systems that transmit and receive in a set direction, the passive system must listen to all angles at all times in order to determine bearing, and this requires a very wide beamwidth. At the same time, a narrow beamwidth is required for locating the source and rejecting ambient noise, thus increasing the signal-to-noise ratio.[xi]

The signal received at the hydrophone array is made up of different sources with multiple frequencies, bandwidths, and strengths located at various bearings and ranges. These sources can include biologics, remote shipping, and sea state noises as well as the radiated signal from the target. Collectively, all these sources make up the background noise in which the beamformer must be able to pick out the target.

To accomplish this, the signal-to-noise ratio must be increased by using time delays and summation techniques. To increase the signal-to-noise ratio, the output of each individual hydrophone is sampled at a different time with each sample of the signal adding constructively, whereas the noise component, being random, adds destructively.

The passive beamforming processor applies a unique set of time delays to the signal to create a particular beam. This process is repeated several times, each with a different set of time delays, in order to listen to many narrow

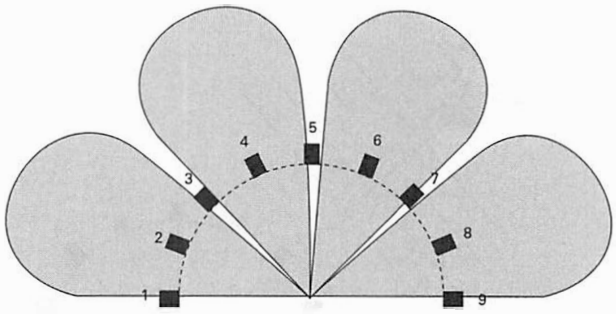


FIGURE 11-18 *Passive beamforming*

beams nearly simultaneously. The result is a set of beams that cover the field of view of the array; see figure 11-18.[xii]

The direction of the signal is then determined by the time delays that maximize the signal-to-noise ratio. The beams do not come from each individual hydrophone but from the combination of sampling from the beamformer processor's forming very narrow beamwidths from the full aperture of the array. To control sidelobes, the beamformer also weights the output of individual hydrophones.

Broadband Display

The output of the beamforming processor is displayed as a bearing time history (BTH), as shown in figure 11-19. The newest information is usually at the top of the display. The beamwidth of the system determines how accurately the bearing can be measured on the display. A common beamwidth is about 5°.

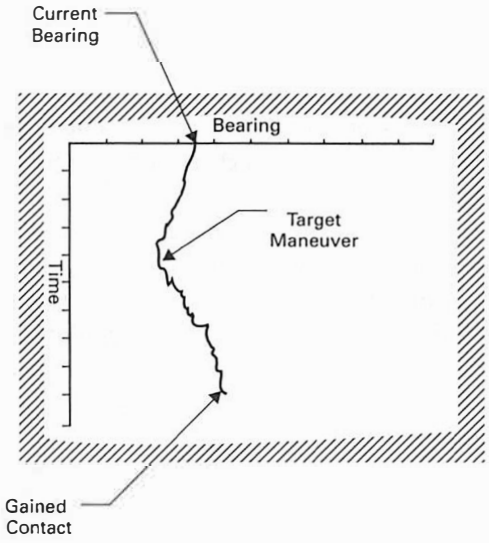


FIGURE 11-19 *Bearing time history (BTH) display*

The total amount of time displayed from top to bottom can be controlled to some extent. A quickly updating display that keeps information for only a few minutes would be useful for close contacts whose bearings are changing rapidly. On the other hand, a long time history is more useful for detecting long-range contacts, whose bearings are only changing slowly.[xiii]

Frequency Analyzer

The frequency analyzer divides the signal into separate frequencies—that is, the spectrum of the signal. For processing purposes, the frequencies are separated into small bands known as frequency bins; see figure 11-20. The width of each bin is called the analysis bandwidth.[xiv]

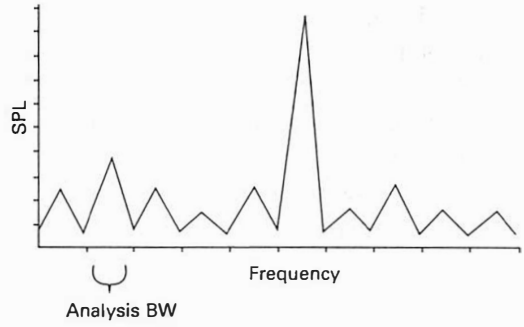


FIGURE 11-20 *Frequency analysis*

Sonar systems gain considerable signal-to-noise improvements by matching the analysis bandwidth to the bandwidth of narrowband sources. If the analysis bandwidth is too wide, then noise from the part of the spectrum beyond the signal is received and the SNR is degraded. If the analysis bandwidth is too narrow, then part of the signal is excluded, also reducing the SNR. The ideal situation occurs when the analysis bandwidth exactly matches the signal bandwidth. This is possible when the characteristics of the signal are well known, which they are for most targets.[xv]

The frequency analyzer separates or "filters" the signal into discrete bins, inside of which the SNR is maximized. The frequency content of the signals from a target provides vital information about the signals' identity and operation. These frequencies are also subject to Doppler shift, just like radar, and therefore can give information about the range rate. However, this requires that the original frequency be known exactly, which is generally not the case. Nonetheless, many important facts can be inferred from the changes in received frequency over time.[xvi]

Narrowband Display

The narrowband display is formed by continuously plotting the time history of the frequency of a particular beam. As time is added to the top of the display, the entire plot moves downward. Hence, this type of presentation is often called a waterfall display; see figure 11-21.

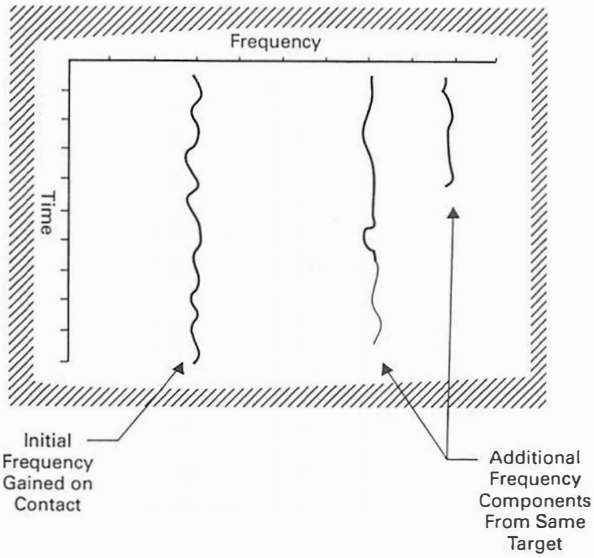


FIGURE 11-21 *Waterfall display*

One useful way in which a narrowband system can be made to search is by having it display several beams simultaneously, each showing a mini-waterfall display; see figure 11-22. This type of detection system is called *grams*. Grams are very useful but require significant operator training and skill because the amount of information displayed increases. Some systems also require the operator to search the entire field of view at a rate of a few beams at a time.

Variable Depth Sonar

To overcome the difficulty of hunting or listening for submarines that take advantage of hiding in shadow zones, a variable depth sonar (VDS) may be used. In this application, a streamlined body containing transducers is towed behind a ship, connected electrically to display and control equipment on the ship. In conjunction with the speed of the ship and scope of the cable, controllable vanes and depth sensors permit the body to be towed at any desired depth.

This system has many advantages in detecting submerged objects. It can operate independently or in combination with hull-mounted sonar. VDS is able to physically penetrate below the sonic layer depth and ensonify the shadow zone, exploiting the sound channel. Also, a deep transducer is not affected by the environmental conditions that limit hull-mounted sonar—that is, quenching and self-noise are greatly reduced. A VDS that operates below a layer is also valuable in hunting for mines. The final advantage is that VDS provides 360° coverage by placing the transducer below the hull baffles, thereby providing detection capability aft of the ship.

The dipping sonar from a helicopter is a variation of a VDS system. In this application, from a hover the helicopter lowers an active transducer by way of a cable to the desired depth. Self-noise is significantly lower because the transducer generates little noise and is isolated from the helicopter-generated noise. Additionally, the search platform—the helicopter—is less vulnerable to a submarine compared to a ship.

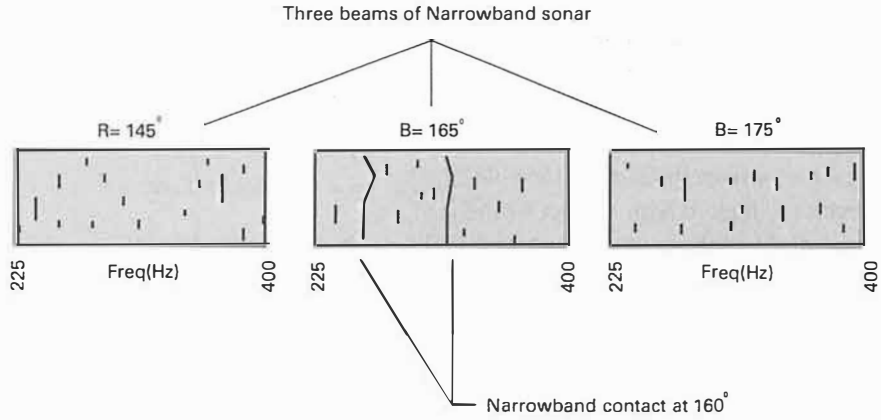


FIGURE 11-22 *Narrowband grams*

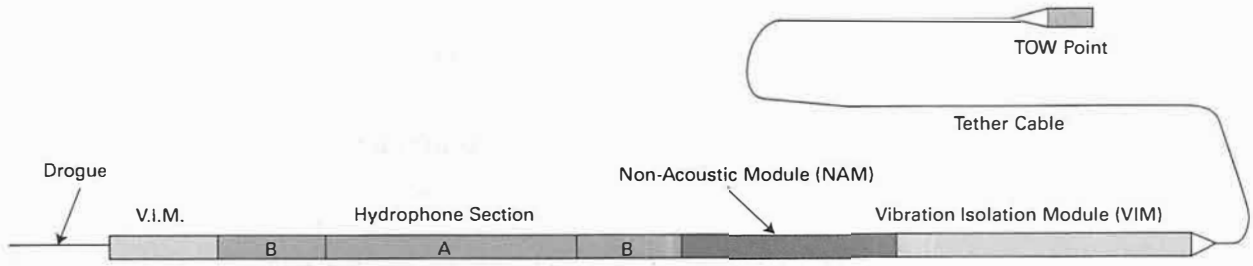


FIGURE 11-23 *Towed array sonar by sections*

Towed Array Sonar System (TASS)

A towed array is a passive system comprising a set of linearly spaced hydrophones that is towed behind a ship or submarine. In addition to the hydrophones, the array consists of heading, water temperature, and depth sensors located in the nonacoustic module that is used to enable the towing platform to determine the depth of the array; see figure 11-23. This is especially important when determining the bearing from which a sound is originating.

Just like the passive sonar system previously discussed, the beamformer processor uses time delays between elements to create narrow beams. This ability is directly related to the directivity index (DI) term in the sonar equation and is developed for a line array of sensors as follows.

Consider first that the ambient noise background of the sea arises principally from wave motion, marine creatures, and the radiated noise of ocean shipping. This component of background noise is isotropic, and thus a single omnidirectional hydrophone seeking a direction signal from a target in an isotropic noise field sees noise over an arc of 360°, whereas the signal sought is available on only a single direction or line of bearing. It is apparent that if a number of hydrophones could be arranged and then "steered" to look in a single direction, noise in all other directions would be rejected and the signal-to-noise ratio in any selected direction would be enhanced.

In figure 11-24, let the line E_0–E_5 represent a line of array elements spaced distance d apart. Let lines *a-a* through *f-f* represent the successive positions of a plane wave of sound energy as it strikes the hydrophone elements of the array from an angle θ with respect to the array axis. Let x be the distance the sound wave travels from the time it strikes one array element until the time it strikes the next. Consider that as the sound wave strikes each element, it generates a voltage proportional to the intensity of the incoming signal.

It can be seen that the output voltage of element E_1 will be delayed at the processor relative to E_0 by the length of

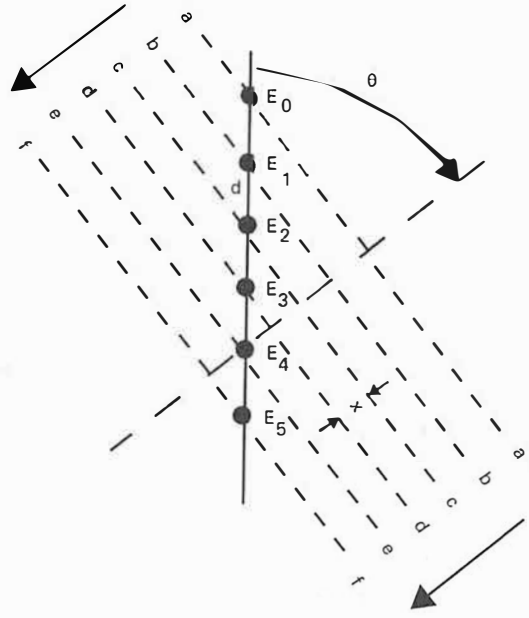


FIGURE 11-24 *Linear array directionality*

time required for the sound to travel the distance x, which will be equal to $(d\cos\theta)/c$. For successive elements, the total delay for an entire array of length l is then given by:

$$\Delta t = \frac{l\cos\theta}{c} \qquad (11\text{-}4)$$

where:

c = sound velocity (m/s)
$l = (n - 1) \cdot d$ (m)
n = number of elements

By electrically delaying the arrival of the signal from each element to the processor according to its position in the array, and then adding these electrical signals, it can be seen that the signal-to-noise ratio for the array input to the processor will be greater than that input by a single hydrophone. This is true because the sinusoidal sound signals added or summed will always be in phase

at all the array elements, whereas the background noise is out-of-phase.

At the same time, the directionality of the array reduces noise arriving from other directions relative to the signal. Note that cosine is used here because the angle θ is measured from the array axis, not from a line normal to that axis as in electronically scanned radars.

This process is used to "steer" electronically the receiving beam of the array. Upon reception of a signal from an unknown direction, the amount of delay among equally spaced elements can be varied until a maximum in-phase signal is detected. This delay can then be used to determine the angle of the target off the array axis.

Note that sound from any target off the "beam" of the array will arrive at all the elements simultaneously with no time delay, and that any sound from targets off the "bow" or "stern" of the array will arrive at successive elements with a maximum time delay. Target positions intermediate to these two will produce delays between zero and the maximum.

These considerations are the basis for the computation of the directivity index in the sonar equation. For the simple line array illustrated, the DI is found by the following:

$$DI = 10\log\left(\frac{2l}{\lambda}\right) \qquad (11\text{-}5)$$

where:

l = array length (from first to last hydrophone) (m)
λ = wavelength for the frequency concerned (m/s)

Towed arrays have upper and lower frequency limits, based on the physical properties of the array. The lower frequency limit is determined by the total length of the array (l). The limit arises when the wavelength of the source is sufficiently large to cause the DI to approach zero. This occurs when the wavelength $\lambda_{max} = 2l$; then $f_{min} = c/2l$.

The upper frequency limit is dependent on the distance between individual hydrophones. As the wavelength is reduced and approaches the distance between individual hydrophones, beam steering cannot be accomplished. This occurs when the spacing between elements (d) equals the wavelength of the source λ_{min}; therefore $f_{max} = c/d$.

To maximize the effectiveness, towed arrays should be extremely long and have many closely spaced elements. This arrangement allows for very narrow beamwidths in the high frequencies and also allows for detection of very low frequencies. Maximizing low-frequency reception is the most important factor since low-frequency sounds are absorbed less than higher frequencies.

Because the array is linear, there is no vertical or horizontal directionality. This causes two problems. The first occurs when there is bottom bounce propagation; see figure 11-25. In this case, the direction of the source cannot be known without further analysis. The second problem is an ambiguity in relative bearing; see figure 11-26. The linear array cannot distinguish signals on the left from those on the right.

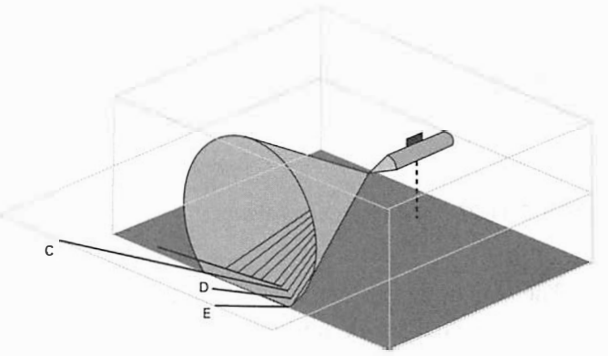


FIGURE 11-25 *Bottom bounce reception*

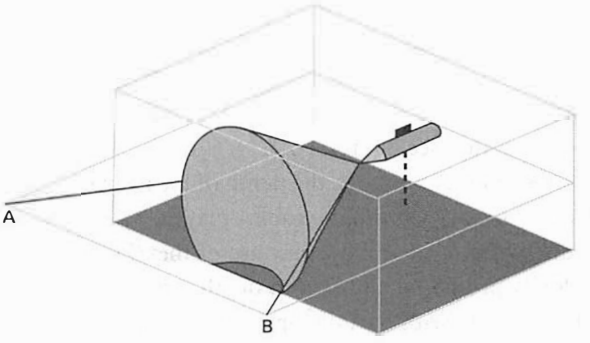


FIGURE 11-26 *Direct path*

Bearing ambiguity can be resolved by maneuvering the ship. When contact is regained, there will again be two ambiguous bearings, but only one of them will match the previous case, assuming that the target hasn't changed bearing in the time interval between legs. Another method to resolve bearing ambiguity is to obtain cross bearings from other platforms.

Sonobuoys

The sonobuoy is an expendable piece of detection equipment that is designed to be dropped from helicopters and patrol-type aircraft and in some instances from ships. The purpose of the sonobuoy is to detect underwater sounds and to transmit these sounds to nearby aircraft or ships through low-powered VHF radio transmissions.

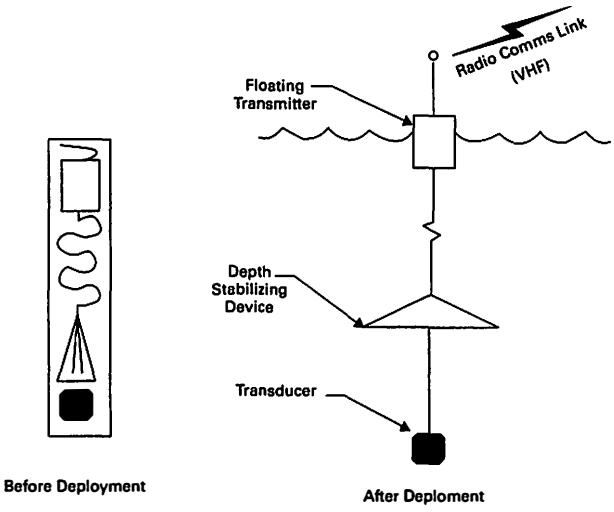


FIGURE 11-27 *Sonobuoy deployment*

The sonobuoy is a receiver/transmitter combination designed to float in an upright position in the water. Upon being deployed from an aircraft, the sonobuoy falls through the air in a vertical position that is stabilized by a parachute. Upon striking the water, the parachute is ejected, a saltwater battery is activated, a small transmitting antenna erects, and a hydrophone or transducer sinks to a predetermined depth. The "buoy" becomes operational usually within 30 to 90 seconds upon entering the water; see figure 11-27.

On some sonobuoys, the depth of the hydrophone (or length of the hydrophone cable) may be selected before launch, thus allowing the best use of the available sound velocity profile. In addition to depth settings, sonobuoys are available with various operating lives, at the end of which they are made to sink (scuttle). This "life" will be short (hours) for buoys intended for localization and long (days) in the case of buoys used in large-area search. The scuttling of the sonobuoy is accomplished by use of water-soluble plugs.

Sounds picked up by the sonobuoy are transmitted to the monitoring receiver on the ship or aircraft. By dropping sonobuoys in patterns encompassing large ocean areas, operators can determine location, course, and speed of a submarine.

Sonobuoys may be divided into three major categories: passive, active, and special-purpose; and two subcategories: omnidirectional and directional.

The omnidirectional passive sonobuoy uses hydrophones to detect the sounds of a target submarine. A single omnidirectional sonobuoy cannot determine the position of the target; however, range estimates can be made if acoustic conditions are known.

The directional passive sonobuoy called a DIFAR (direction finding acoustic receiver) provides an output that indicates the bearing of the source relative to the sonobuoy position. This is accomplished by a magnetic compass, which, combined with the hydrophone, determines bearing angle of the target with respect to magnetic north. These inputs are then multiplexed and transmitted to a receiving platform.

The VLAD (vertical thin-line array DIFAR) is a passive sonobuoy that has an array of sonobuoys much like a towed array, arranged vertically versus horizontally. Beamforming technology is then applied, allowing significant noise reduction and steering of the main beam.

Active sonobuoys' pulse initiation and pulse duration are either self-timed or commanded offboard. In current use is the DICASS (directional command activated sonobuoy system), which receives activation commands from the controlling platform via radio link. This type of sonobuoy provides both bearing and range information to the target by activating the onboard receiver during the interval between pulses and by determining directions relative to magnetic north.

In addition to the active and passive sonobuoys, there are two special-purpose types. The first is used to obtain a temperature versus depth profile to allow prediction of sound propagation conditions, similar to what an XBT does. The depth is determined by timing the descent of the temperature probe.

The second special-purpose sonobuoy is designed to act as a communications link between aircraft and submarines.

Multistatic Sonar

Multistatic sonar is a combination of an active system for transmission at one location with one (bistatic) or more (multistatic) passive systems for reception at other locations. There are several advantages to this type of sonar operation. The receiving system does not suffer from the same reverberation limitations as the active system. Thus, source levels can be very high, approaching the quenching limit. In fact, some systems use explosive-driven projectors having no quenching limit and thus have extremely large source levels.

Additionally, the transmission loss is less than that of the full two-way loss of a normal active system; see figure 11-28. Therefore, the sound source can be very far away, outside the sensor or weapon range of the target. To obtain accurate ranges, however, a precise time of the active signal and the exact position of both the source and the receivers must be known.

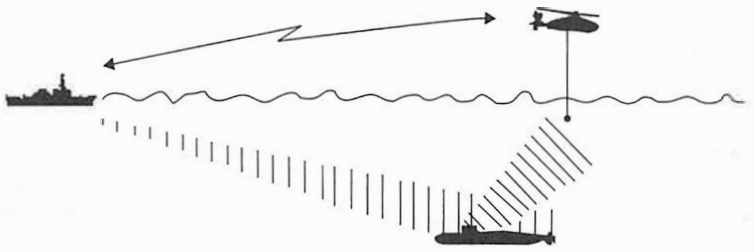


FIGURE 11-28 *Bistatic sonar*

Nonacoustic Detection

Visual

Submarines at or near the ocean surface are extremely vulnerable to visual detection. Periscopes, snorkeling masts, and antennas leave significant visual signs that low-flying aircraft are able to detect up to 10 miles away; see figure 11-29.

Protrusions above the water when submarines travel above 4 knots leave behind a white wake, called a feather, up to a few yards long. The remnant of a feather is called a scar. A scar is formed by bubbles and foam from the feather and can reach up to 10 yards in length.

A submerged submarine's shadow when at shallow depths is also detectable by aircraft where the ocean bottom is lightly colored. Phosphorescence, the light produced by biological organisms when disturbed, also gives away a submarine's position at night..

Radar

Radars can be specifically designed to detect the very small radar cross sections of periscopes and masts. These radars have very precise range resolution and narrow beamwidths for good bearing resolution. The antennas are also vertically polarized to match the target's orientation. Inverse synthetic aperture radar is a specific system that has proved capable of detecting periscopes and masts.

A type of radar that measures the height of the ocean is also able to detect the presence of a submarine. When a shallow submarine is moving quickly, a phenomenon known as a Bernoulli hump is formed. A Bernoulli hump is the rising of the sea surface caused by the submarine's motion. Because of the extensive signal processing involved with this system (for example, taking out the random noise of the waves), analysis takes several hours to complete. Thus this method provides nonreal time information but does tell of a submarine passing.

Infrared Detection

Diesel-electric submarines must periodically use their internal combustion engines to recharge electric batteries. When a submarine is recharging batteries while submerged, exhaust is discharged to the surface through a device called a snorkel. The exhaust gases from these engines are a detectable source in the infrared portion of the electromagnetic spectrum; they are, however, of limited utility because the gases are often mixed with a fine mist of water. Additionally, opportunities for detection are limited because at slow speeds, diesel-electric submarines snorkel for only a few hours each day.

Laser

The use of lasers for the detection of shallow underwater objects is yet another method of nonacoustic undersea warfare. Because of the absorption characteristics of electromagnetic energy in water, the best wavelength to penetrate to a useful depth is between 0.2μm and 0.4μm, corresponding to the visible light color bluish green.



FIGURE 11-29 *Swirl behind a submarine after a dive*[xvii]

The laser detection systems often deployed from a helicopter are mounted on an outboard pod. The laser is emitted in pulse mode and continuously scans swaths of area perpendicular to the flight path of the aircraft.

The system pulses the laser into the water, and a synchronizer signals camera shutters to open, allowing the laser reflections from an object at a selected depth. When multiple cameras are used, multiple depths can be scanned on a single pulse. The received images are then processed to obtain a three-dimentional picture of the water column that displays objects the size of sea mines to a depth of 30 meters.

Magnetic Anomaly Detection (MAD)

Another method of detecting a submerged submarine is through the use of magnetic anomaly detection (MAD), which uses the principle that metallic objects disturb the magnetic lines of force of the earth. Light, radar, or sound energy cannot pass from air into water and return to the air in any degree that is usable for airborne detection. However, the lines of force in a magnetic field are able to make this transition between media almost undisturbed, because magnetic lines of force pass through both water and air in similar manners. Consequently, a submarine beneath the ocean's surface, which causes a distortion or anomaly in the earth's magnetic field, can be detected from a position in the air above the submarine. The detection of this anomaly is the essential function of MAD equipment.

When a ship or submarine hull is being fabricated, it is subjected to heat (welding) and to impact (riveting). Ferrous metal contains groups of iron molecules called domains. Each domain is a tiny magnet that has its own magnetic field with a north and a south pole. When the domains are not aligned along any axis, but point in different directions at random, there is a negligible magnetic pattern. However, if the metal is put into a constant magnetic field and its particles are agitated, as they would be through hammering or heating, the domains tend to orient themselves so that their north poles point toward the south pole of the field, and conversely so that their south poles point toward the north pole of the field.

All the fields of the domains then have an additive effect, and a piece of ferrous metal so treated has a magnetic field of its own. Although the earth's magnetic field is not strong, a ship's hull contains so much steel that it acquires a significant and permanent magnetic field during construction. A ship's magnetic field has three main components—vertical, longitudinal, and athwartship, the sum total of which is comprised in the complete magnetic field, as shown in figure 11-30.

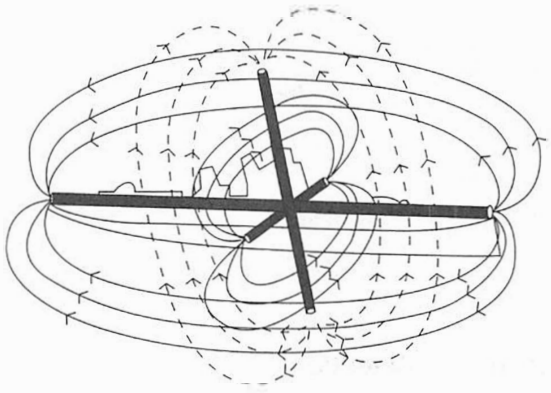


FIGURE 11-30 *Permanent magnetic flux lines about a ship*

The steel in a ship also has the effect of causing the earth's lines of force (flux) to move out of their normal positions and be concentrated at the ship. This is called the induced field, and varies with the heading of the ship.

A ship's total magnetic field or "magnetic signature" at any point on the earth's surface is a combination of its permanent and induced magnetic fields. A ship's magnetic field may be reduced substantially by using degaussing coils, often in conjunction with the process of "deperming," or neutralizing the permanent magnetism of a ship; but for practical purposes, it is not possible to eliminate such fields entirely.

The lines comprising the earth's natural magnetic field do not always run straight north and south. If traced along a typical 200-kilometer path, the field twists at places to the east or west and assumes different angles with the horizontal. Changes in the east-west direction are known as angles of variation, whereas the angle between the lines of force and the horizontal is known as the angle of dip. Short-trace variation and a dip in the area of a large mass of ferrous material, although extremely minute, are measurable with a sensitive magnetometer; see figure 11-31.

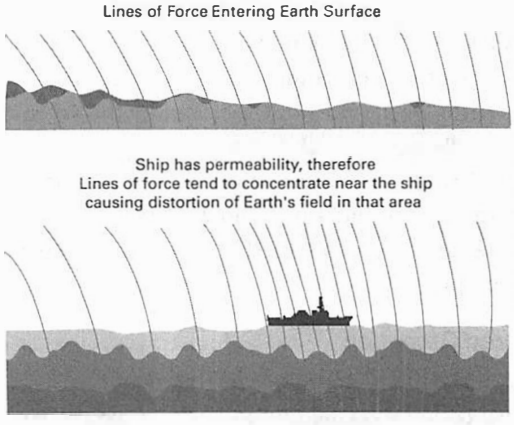


FIGURE 11-31 *Effect of ship's hull on the earth's magnetic field*

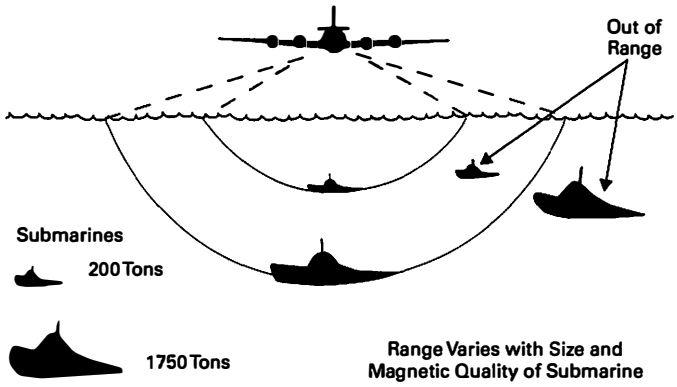


FIGURE 11-32 *Magnetic anomaly detection (MAD) sweep*

The function, then, of airborne MAD equipment is to detect the submarine-caused anomaly in the earth's magnetic field. Slant detection ranges are on the order of 500 meters from the sensor. The depth at which a submarine can be detected is a function both of the size of the submarine and of how close the sensor is flown to the surface of the water; see figure 11-32.

Improving MAD detection ranges has proved extremely difficult. Increasing the sensitivity of the MAD gear is technically feasible, but, operationally, due to the nature of the magnetic anomaly, it is not productive. The magnetic field of a source such as a submarine, falls off as the third power of the distance; hence, an eightfold sensitivity increase would serve merely to double the range.

Additionally, magnetometers are nondirectional, and the physics of magnetic fields do not permit the building of instruments that would respond preferentially to a field coming from a particular direction. Hence, a valid submarine-caused disturbance frequently is masked by spurious noise.

Also, the ocean floor in many areas contains magnetic ore bodies and similar formations of rock, which can confuse the signal. Further confusion is caused by magnetic storms, which produce small, but significant, variations in the earth's field.

MAD equipment is primarily used by aircraft as a localization/targeting sensor. A helicopter's small turn radius allows for optimum MAD employment. Helicopters tow the sensor on a 25-to-55 meter cable behind and below the aircraft, which reduces helicopter self-noise. Fixed-wing ASW aircraft MAD configurations are provided in an aircraft tail boom. Because of the relatively short detection ranges possible, MAD is not generally utilized as an initial detection sensor.[xviii]

Fathometer

Water depth can be measured in several ways. One method is to drop a weighted, distance-marked line to the bottom and observe the depth directly from the line. The chief disadvantages of this method are that its use are limited to very shallow water and that operation is slow.

The use of sound is the more common method of measuring water depth. A sound pulse, directed toward the bottom, is transmitted, and its echo is received. The time between pulse transmission and echo reception is measured and, based on the speed of sound in water, the depth is determined.

Such a device is known as fathometer. It is a navigational instrument that operates on acoustic principles rather than by radar or visual means. Its use is especially critical when following bathymetric navigation methods. In addition, if the system is aimed upward, submarines can navigate under ice.

Underwater Communications

For many reasons it may be necessary for ships, aircraft, and submerged submarines to communicate with each other. Of paramount importance is the safety of the submarine and its crew.

During exercises, the ship can advise the submarine when it is safe to surface. Should an emergency arise aboard the submarine, the ship can be so informed. Exercises can be started and stopped, or an exercise in progress can be modified. Attack accuracy can be signaled by the submarine to the attacking ship.

A number of devices exist that facilitate these communications, all of which incorporate either one-or two-way voice or tone/sound signal generation that utilizes sonar or sonobuoy-type equipment.

The range and quality of transmission vary with water conditions, local noise level, and reverberation effects. Under optimum sonar conditions, however, communication between ships should be possible at ranges out to 12,000 meters.

Under the same conditions, submarines achieve a greater range. If the submarines are operating in a sound channel, the communication range may be many kilometers greater than that achieved by ships. Local noise—caused by a ship's movement through the water, machinery, screws, and so on—can reduce the range to less than half the normal range. Severe reverberation effects may also cause a significant reduction.

Because of the characteristics of underwater sound transmission, the amount of data that can be communicated through sonic means is severely limited. The primary and most reliable means of communication with submarines and other forces is through various radio systems that range from extremely low frequency (ELF) to ultrahigh frequency (UHF) satellite. The major disadvantage, however, is the requirement to be close to the surface to facilitate the use of various antennas during communication periods. Because of this, during tactical operations, the only viable form of emergent communications to a submarine is through acoustic signals.

Subsurface communication in the electro-optic spectrum has generated much effort in research and development. Future systems may utilize laser communications, allowing short-duration submerged exchanges.

Countermeasures

Early sonar operators rapidly discovered that they could passively detect active acoustic devices at ranges greater than the range at which the active devices could detect the passive platform. Tacticians reasoned that this counterdetection situation could be exploited much the same as electronic surveillance and radar.

The initial underwater threat detection and evaluation capability employed the listening capability of the installed attack sonar set. This was limited to the bandwidth of the sonar receiver until the development of separate, interconnected hydrophone arrays for active sonars. By the 1950s, dedicated underwater intercept receivers such as the AN/WLR-2 had been developed.

Acoustic countermeasures began with the Pillenwerfer used by German U-boats when the Allied hunter-killer groups began to take their toll in late 1943 and early 1944. This device, composed of lithium hydride, acted like a giant Alka-Seltzer tablet, creating thousands of gas bubbles that returned a solid echo similar to the echo of a submarine. This was reasonably effective as long as the attacking units didn't see the submarine and the decoy at the same time. Because the decoy did not cause a Doppler shift, a trained operator could tell the difference.

Dealing with acoustic countermeasures requires the same basic approach as electronic protection. Again, the best countermeasure is an experienced and well-trained operator; nevertheless, improvements such as computer signal processing, various filters and delay lines, and automated Doppler detection capability are of great help.

Acoustic Intercept Receivers

Initial attempts at producing underwater intercept receivers were less than satisfactory because the equipment had a high false-alarm rate due to ambient noise. Separate receivers were required for initial intercept and for determination of the bearing of the noise source. Later receivers employed a triangular-shaped array of three hydrophones that determined bearing by measuring the time difference between the arrival of the signal at each hydrophone.

Noise reduction was accomplished by using delay lines and filters that allowed only signals longer than the delay to reach the display. Short noise pulses resulted in no output. A receiver designed in this way could display frequency and bearing simultaneously after receiving one ping.

With the development of active acoustic homing torpedoes, this type of response is required of all acoustic intercept receivers in order for there to be time to employ torpedo evasion tactics. Modern receivers categorize intercepts and assign priorities according to the potential threat. In this situation, torpedoes would have the highest priority, with other intercepts having lesser priority, the least being search sonars in a long-range mode.

Acoustic Countermeasures

The initial research of acoustic countermeasures focused on the control of active emissions from sonars and the reduction of self-noise. As with electronic emissions, acoustic emissions should be kept to a minimum and, if possible, used only after counterdetection has occurred. However, if analysis of the FOM problem indicates that passive detection of an enemy is not likely, then active detection may be the only option.

Active operation must take into account the tactical situation. Additionally, transmission power levels must be minimized as much as possible to prevent counterdetection. During active sonar prosecutions, a change to

short-range mode will shorten the time between pings and let an adversary know that the undersea warfare platform is in contact. This could result in subsequent target maneuvering and an unsuccessful attack, with no second opportunity.

Countermeasures for use against underwater acoustic sensors, such as homing devices and fuzes, developed as these systems evolved. A mine, activated by the sounds emanating from a power-driven ship underway, was developed by the British in the late stages of World War I. However, the first operational use of acoustic mines was accomplished by the German navy against the British in the fall of 1940. Work started immediately to develop a noisemaker that would project noise far enough ahead of the ship to detonate the mine before the ship was close enough to be damaged by the explosion.

Among the earliest methods developed to counter the acoustic mine were explosives, compressed air, electrically operated hammers striking the inside of a watertight case suspended over the side of the ship, and parallel pipe devices that were towed through the water to cause the pipes to bang together as the pressure between the pipes was reduced by the faster motion of the water through a constrained space (known as Bernoulli's principle).

There is a considerable similarity between the early attempts and the present-day U.S. Navy acoustic sweeps. Early sweep gear consisted of vanes or propellers attached to noisemakers not too different from a New Year's Eve whirling ratchet. When towed through the water, the propellers turned and the ratchet device produced noise. Devices similar in principle are still used; for instance, the U.S. Navy's airborne acoustic sweep device depends on a water turbine to produce a turning motion for its noise-making capability.

The early noisemakers were turned on and kept operating at one intensity while the sweeper was in the field. Subsequently, controllers were added that enabled the intensity of the output to be varied to simulate the approach of a ship target. The same approach can be employed as a means of decoying homing torpedoes. There are noisemakers that can be towed from ships or submarines, providing a lucrative alternate target when the torpedo-target geometry is such that there is a wide azimuth separation between the noisemaker and the target.

Modern acoustic towed decoys such as the AN/SLQ-25 NIXIE employ electronic or electromechanical means to produce the required signals. Submarines employ small torpedolike mobile decoys that are equipped with transponders capable of returning a realistic signal to active sonars. With noisemaking capability, they can simulate submarine passive signatures to some degree.

This same technology has been used to provide small expendable training targets, as depicted in figure 11-33, for undersea warfare forces at much less cost than that of a real submarine providing target services. Though small (about the size of a sonobuoy), the transponders in these targets make them look as large as a real submarine to the detecting sonar.

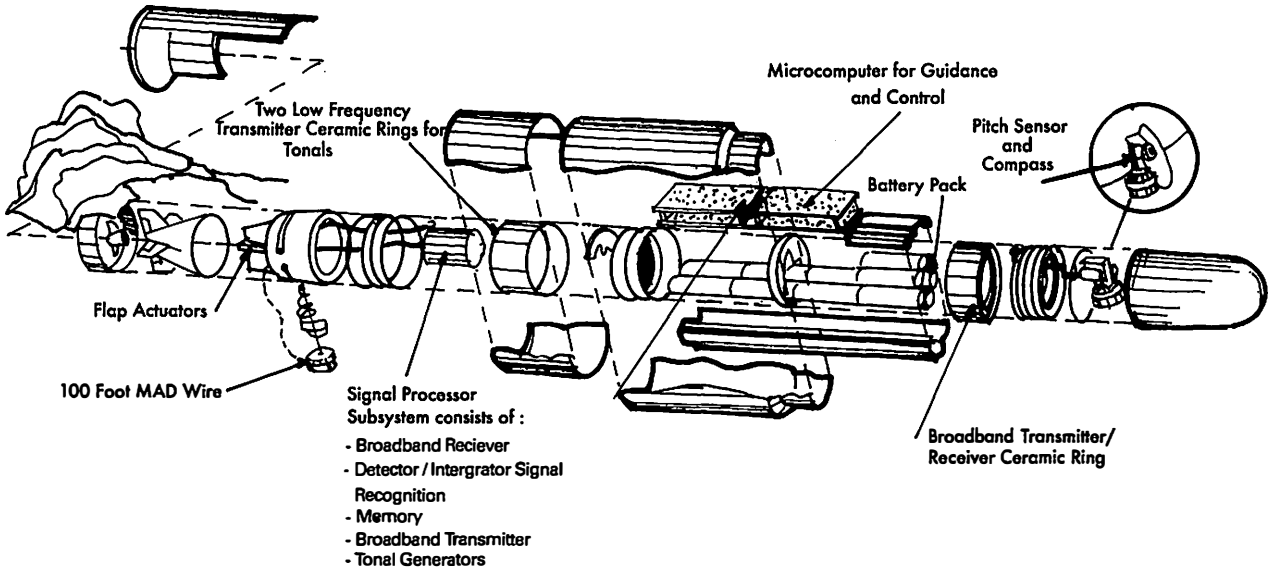


FIGURE 11-33 *Expendable mobile target*

Acoustic jammers that provide random noise or false targets to hostile active sonars are available. With the power of today's sonars, a jammer must have an output source level of at least 200dB (ref. 1μPa) to be effective when mounted aboard a submarine. The rules governing jammer output, "burnthrough," jammer bandwidth versus sonar bandwidth, and sonar frequency agility are similar to those dealing with radar and electronic attack devices.

Air bubbles can be employed to mask potential targets or to provide alternate targets. The large difference in characteristic impedance between the air bubbles and the surrounding water makes them very efficient reflectors of acoustic energy. Very little sound will penetrate a curtain of air bubbles, also making them very efficient at masking noise sources. This is the basis of the belt masker system installed aboard surface combatants.

---

[i] Grace, 63
[ii] Urick, 31
[iii] Grace, 64
[iv] U.S. Navy, *Principles*, 140
[v] Kinsler, 180
[vi] Grace, 94
[vii] Urick, 64
[viii] Hall, 214
[ix] Urick, 322
[x] Grace, 98
[xi] Hall, 216
[xii] Hall, 217
[xiii] Hall, 218
[xiv] Hall, 218
[xv] Hall, 219
[xvi] Hall, 219
[xvii] U.S. Navy, *Principles*, 117
[xviii] *FAS*, chap. 1

12 Introduction to Automatic Control Systems

Introduction

A control system is an arrangement of physical components connected or related in such a manner as to command, direct, or regulate itself or another system.[i] In simpler terms, it is a means of regulating a process.

Control systems have been around for centuries and currently surround our everyday lives. The human body itself is a control system maintaining body temperature at approximately 98.6°F. When a control system is self-regulating, which means it needs no outside intervention, it is deemed an automatic control system. Maintaining the speed in your car is accomplished via a control system: your foot on the pedal. When you activate the cruise control, you are using an automatic control system.

An essential part of any automatic control system is feedback, which is some sort of signal that is returned or "fed back" to the control device to achieve a desired output.

A Short History of Control Systems

Control systems are not a new concept. More than 2,000 years ago, the Greeks developed a mechanism to maintain a constant water level in a tank, much akin to what is used in today's flush toilets.[ii] Others during that time period developed a system that would automatically maintain the level of fuel in an oil lamp. Both these systems were position controls and had position feedback.

Speed control was one of the first tasks pursued in the field of automatic control engineering. With the advent of the steam engine in the 1700s, it was desirable to maintain the speed of the engines. James Watt is credited with creating the fly ball governor as shown in figure 12-1. When the engine slowed due to higher loading, the ball weights would hang at a smaller angle relative to the axis of rotation. This would push a sleeve connected to actuators that would open the butterfly valve, allowing more steam into the engine. Adjusting the length of the actuator arms was required if one desired to make changes in the actual controlled speed.[iii]

The twentieth century saw an increase in the use of control systems. During the 1920s, Elimen Sperry developed an automatic steering mechanism for ships. As scientists delved into the subject of control systems, improvements were made beyond the simple proportional control of their predecessors and added integral and derivative control.

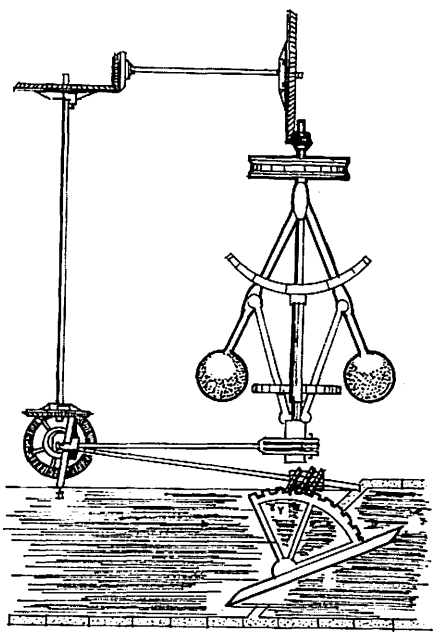


FIGURE 12-1 *Watt's fly ball governor*

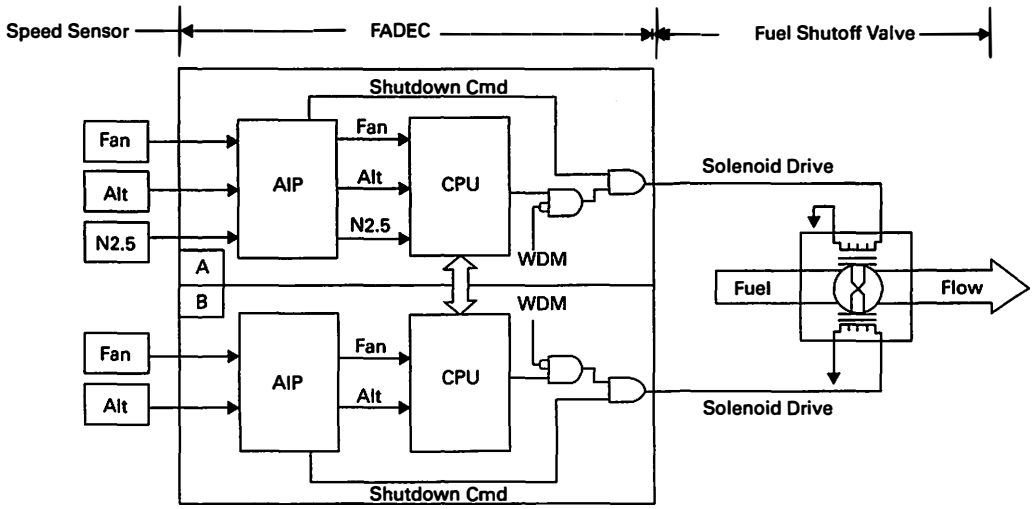


FIGURE 12-2 *Control system diagram for a full authority digital electronic control (FADEC)*[iv]

Among these improvements were servomechanisms and power-amplifying systems. In general, these technologies translate a small input into a greater force to position a specific component. The training of a shipboard gun mount is an example of this. A servomechanism positions the gun mount based on the input from the operator, and an automatic control system stabilizes the gun for ship motion. Other examples of servomechanisms are the power steering and power brakes found on most automobiles.

The modern world is surrounded by control systems. On board a ship, the rudder is controlled via a servomechanism, temperatures are maintained via an environmental control system, and the propulsion plant is controlled and optimized via combustion control systems. Near and dear to many is the CD or DVD player, a device that would not be possible without the precise control enabled by an automatic control system.

Control systems have greatly progressed in complexity. Simple systems like Watt's governor had just one input and fed back one signal in a near-linear relationship. More modern systems handle multiple inputs and feedback signals, including control nonlinearities.

Many modern aircraft—commercial as well as tactical—have digital flight control systems. Such systems are essential to enabling an aircraft to fly in a variety of flight conditions, from takeoff and landing to supersonic speeds. Flight control systems also enable the aircraft to have relaxed stability criteria, which enables increased maneuverability while reducing drag.

Powering these modern aircraft is the gas turbine engine. It is controlled by a complex control system, a portion of which is depicted in figure 12-2. This system has at least ten inputs that are used to calculate the fuel flow. The control system ensures proper functioning of this engine enabling the engine to operate from a dead stop, through a carrier catapult launch, and on to supersonic speeds.

This is accomplished with little or no human input. Without such complex control systems, the gas turbine engine would be unable to attain the level of performance that it has.

In examining any control system, a key analysis is how it responds to an input as a function of time. This is known as the system's time response. The time response is a graph of the amplitude of the system's output versus time, and describes how long the output takes to get to the desired value while showing the transient it may undergo to get there.

Sometimes the output never reaches the desired value, resulting in a "steady state error." Sometimes the output reaches the desired value but takes too long to get there. Other times, the output reaches the desired value, but only after oscillating radically about the desired value before settling down. Still other times, a system may exhibit "unstable operation" where it never converges to the desired value.

Modeling

One of the first steps in designing a control system is modeling, which transforms the physical components of

the system, known as the plant, into mathematical terms. More specifically, the plant is represented by a differential equation—that is, an equation containing at least one derivative. The plant could be any number of things—a simple motor or a complex space shuttle.

Accurate modeling is essential to developing a reliable control system. Modeling is also important for simulations. Mathematical models enable us to make trials, evaluations, and predictions without the actual expense of building a physical system. Categories of systems that are modeled using the same techniques include thermal, fluid, mechanical, and electrical.

Mechanical Systems

There are two basic types of mechanical systems, translational and rotational. An example of a translational system is shown in figure 12-3. Note how a force, F, is applied as an input and how a displacement, X, results as an output (or response). Also present are a spring, damper, and mass, which will affect how the system responds, depending on the respective values of K, B, or M.

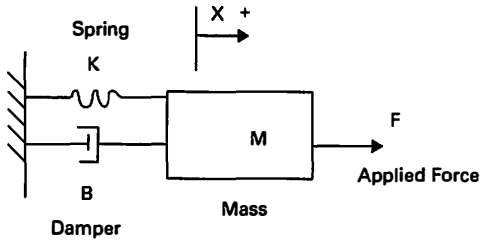


FIGURE 12-3 *A simple translational system*

An example of a rotational system is shown in figure 12-4. A rotational force, known as a torque, T, is applied as an input while a rotational displacement, θ, results as an output. Several system components are present that will affect the way the system responds to the applied torque.

This system consists of a rotational spring as well as two components with rotational damping effects, a hydraulic coupler and a bearing. Also present is the system's moment of inertia, J, which is analogous to mass but is in the rotational vice translational direction.

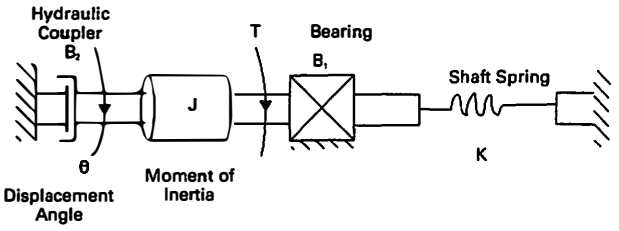


FIGURE 12-4 *A simple rotational system*

Translational Systems

A differential equation can be developed for a simple translational system. The example in figure 12-5 is a system that absorbs the recoil force of a gun firing. The gun will have some mass, M; a shock absorber/damper, B; and a spring, K. The force applied, F, represents the force due to gun recoil.

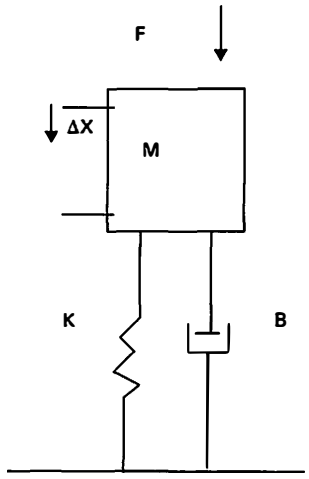


FIGURE 12-5 *Model of a gun recoil system*

Physics equations show that the sum of the forces equals the mass times the acceleration.

$$\sum f = Ma$$

$$\sum f = F_{applied} - F_{spring} - F_{damper}$$

$$\sum f = F_{applied} - Kx - Bv$$

$$Ma = F_{applied} - Kx - Bv$$

Also, recall that velocity is the rate of change (derivative) of position while acceleration is the rate of change (derivative) of velocity. Acceleration is therefore the second derivative of position. The resulting equation is:

$$M\ddot{x} = F - Kx - B\dot{x} \qquad (12\text{-}1)$$

where

$\ddot{x}$ = acceleration
$\dot{x}$ = velocity
x = displacement

Rearranging,

$$F = M\ddot{x} + B\dot{x} + Kx \qquad (12\text{-}2)$$

The previous is a basic differential equation for a translational system with an input force, F, and an output displacement, X.

Rotational Systems

A differential equation can also be developed for a rotational system—that is, a system that rotates about an axis. Continuing with the gun mount example from a rotational perspective, the gun has a mass moment of inertia, J, a coefficient of damping due to couplings and bearings, B, and a rotational spring with stiffness, K. A picture of the gun mount is shown in figure 12-6, and a model is shown in figure 12-7.

$$\sum \text{torque} = J\ddot{\theta}$$

$$\sum \text{torque} = T - T_{\text{spring}} - T_{\text{damper}}$$

$$\sum \text{torque} = T - K\theta - B\dot{\theta}$$

$$J\ddot{\theta} = T - K\theta - B\dot{\theta}$$

where

$\ddot{\theta}$ = angular acceleration
$\dot{\theta}$ = angular velocity
θ = angular position

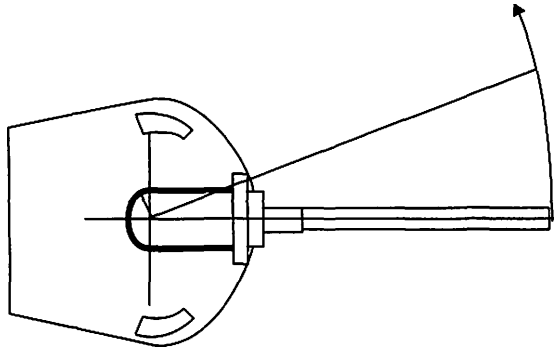


FIGURE 12-6 *Overhead view of a gun mount*[v]

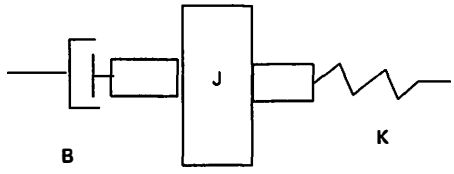


FIGURE 12-7 *Model of a rotational gun mount system*

Rearranging,

$$T = J\ddot{\theta} + B\dot{\theta} + K\theta \qquad (12\text{-}3)$$

Similarly, if the gun mount's response to a torque in the elevation direction were analyzed, the following differential equation would result:

$$T_e = J_e\ddot{\Phi} + B_e\dot{\Phi} + K_e\Phi \qquad (12\text{-}4)$$

Equations 12-3 and 12-4 represent differential equations for rotational systems with an input torque, T, and an output displacement, θ.

Electrical Systems

Electrical systems or circuits can also be modeled in a manner similar to mechanical systems. Instead of force or torque as an input, electrical systems use voltage as the input. The output or response for these systems is current. The typical circuits of an electrical system that will affect the system's response are resistors, capacitors, and inductors.

Figure 12-8 shows a typical electrical circuit with resistance, R, capacitance, C, and inductance, L.

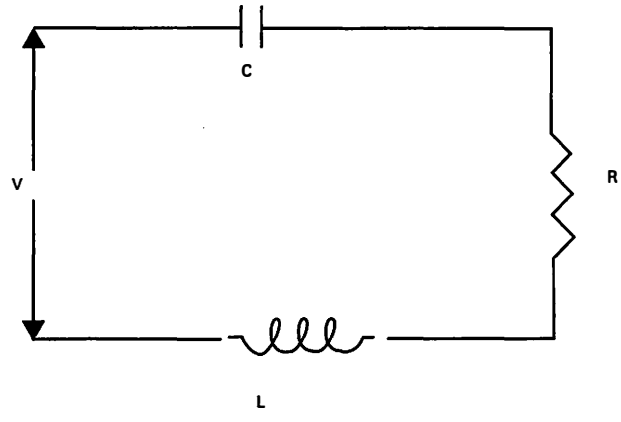


FIGURE 12-8 *A simple electrical circuit*

Using Kirchhoff's voltage law, the input voltage is equal to the sum of the voltage drops across the components in the circuit:

$$V_a = V_{\text{capacitor}} + V_{\text{resistor}} + V_{\text{inductor}} \qquad (12\text{-}5)$$

The voltage drop across a resistor is:

$$V_{\text{resistor}} = iR$$

The voltage drop across a capacitor is:

$$V_{\text{capacitor}} = \frac{1}{C}\int i \cdot dt$$

The voltage drop across an inductor is:

$$V_{\text{inductor}} = L\frac{di}{dt}$$

Equation 12-5 becomes:

$$V_a = iR + L\frac{di}{dt} + \frac{1}{C}\int i \cdot dt \qquad (12\text{-}6)$$

Equation 12-6 is a differential equation for an electrical circuit with input voltage V_a and output current i.

Laplace Transforms

An important part of the mathematical modeling of systems is the use of the Laplace transform. The Laplace transform is a mathematical function that shifts an equation from the "time domain" to the "Laplace" or "s domain." In other words, the equation is no longer a function of the variable, t, but instead becomes a function of the variable, s. The exact mathematical definition of the Laplace transform for any function $f(t)$ is as follows:

$$L[f(t)] = \int_0^\infty e^{-st} f(t)dt = F(s) \qquad (12\text{-}7)$$

where L is the Laplace operator, and the new function in the s domain is delineated by a capital letter F—that is, $F(s)$.

Shifting a function to the Laplace domain has several advantages for mathematical modeling. First, it enables the use of algebra instead of complex calculus functions as a means of solving differential equations. Second, it greatly facilitates mathematical modeling by allowing each subsystem within a system to be represented by a separate mathematical function called a transfer function. Finally, the use of the Laplace domain is consistent with many mathematical modeling software programs such as MATLAB.

This text will not derive each Laplace transform, but will instead reference the table of Laplace transforms to convert from the time domain to the s domain and vice versa. Table 12-1 is a table of Laplace transforms necessary for the student.

Recall equation 12-2:

$$F = M\ddot{x} + B\dot{x} + Kx$$

This differential equation, which is a function of time, can be shifted to the Laplace domain using table 12-1 as follows:

$$F(s) = Ms^2X(s) + BsX(s) + KX(s) \qquad (12\text{-}8)$$

Note how the first derivative in the Laplace domain is represented by multiplying by s, and the second derivative is represented by multiplying by s^2.

TABLE 12-1 *Simplified Laplace transform pairs*

Any Time Domain Function ($f(t)$, $x(t)$, $\tau(t)$, $\theta(t)$, $v(t)$, $i(t)$)	Laplace Domain Functions ($F(s) = L[f(t)]$) ($F(s)$, $X(s)$, $T(s)$, $\Theta(s)$, $V(s)$, $I(s)$)
$f(t)$	$\int_0^\infty f(t)e^{-st}dt = F(s)$
$x(t) + y(t)$	$X(s) + Y(s)$
$nf(t)$ where n is any integer	$nF(s)$ where n is any integer
$\dot{f}(t)$	$sF(s)$
$\ddot{f}(t)$	$s^2F(s)$
$\int_0^t f(t)dt$	$\frac{1}{s}F(s)$
Any integer n (step function)	$\frac{n}{s}$
nt where n is any integer	$\frac{n}{s^2}$
nt^2 where n is any integer	$\frac{2n}{s^3}$
$\sin(\omega t)$	$\frac{\omega}{s^2+\omega^2}$
$\cos(\omega t)$	$\frac{s}{s^2+\omega^2}$

Solving for $X(s)$:

$$X(s) = \frac{F(s)}{Ms^2 + Bs + K} \qquad (12\text{-}9)$$

At this point, the differential equation can be solved by converting the function $X(s)$ back to the time domain—$x(t)$—by use of the inverse Laplace transform.

Transfer Functions

Any system can be represented by a box with an input and an output as shown in figure 12-9:

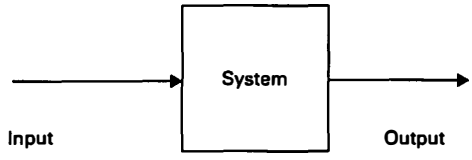


FIGURE 12-9 *System block diagram*

The mathematical function describing the system is known as the transfer function, and is usually given the symbol $G(s)$. The system's transfer function is also

expressed as a ratio of the system output divided by the system input—that is:

$$G(s)=\frac{\text{output}}{\text{input}}$$

Referring to the translational system example in figure 12-3, the block diagram is shown in figure 12-10.

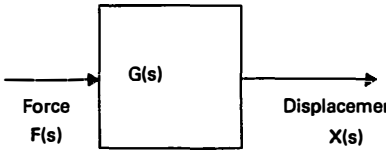


FIGURE 12-10 *Translational system block diagram*

Equation 12-9 can be solved for the ratio of the output, $X(s)$, divided by the input, $F(s)$, as follows:

$$G(s)=\frac{X(s)}{F(s)}=\frac{1}{Ms^2+Bs+K} \qquad (12\text{-}10)$$

The expression in equation 12-10 is the transfer function, $G(s)$, of the translational system of figure 12-3. By convention, the transfer function is properly expressed in "root locus form," which means that the highest power of s in the denominator has a coefficient of 1. Dividing numerator and denominator of the transfer function by the mass, M, gives the root locus expression in equation 12-11.

$$G(s)=\frac{X(s)}{F(s)}=\frac{\frac{1}{M}}{s^2+\frac{B}{M}s+\frac{K}{M}} \qquad (12\text{-}11)$$

Similarly, transfer functions for the rotational system in figure 12-7 and the electrical system in figure 12-8 can be developed. For the rotational system, equation 12-3 is first converted to the Laplace domain as follows:

$$T(s)=Js^2\theta(s)+Bs\theta(s)+K\theta(s) \qquad (12\text{-}12)$$

The system block diagram is shown in figure 12-11.

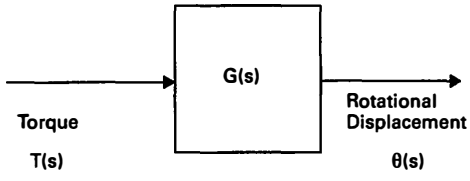


FIGURE 12-11 *Rotational system block diagram*

Equation 12-12 is rearranged, and the rotational displacement, $\theta(s)$, is divided by torque as follows:

$$G(s)=\frac{\Theta(s)}{T(s)}=\frac{1}{Js^2+Bs+K} \qquad (12\text{-}13)$$

Equation 12-13 is then converted to root locus form:

$$G(s)=\frac{\frac{1}{J}}{s^2+\frac{B}{J}s+\frac{K}{J}} \qquad (12\text{-}14)$$

For the electrical system of figure 12-8, equation 12-6 is first converted to the Laplace domain as follows:

$$V_a(s)=RI(s)+LsI(s)+\frac{1}{SC}I(s) \qquad (12\text{-}15)$$

The system block diagram is shown in figure 12-12.

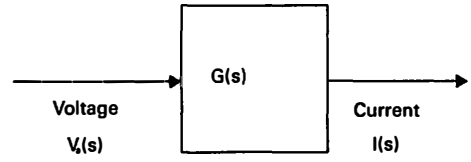


FIGURE 12-12 *Electrical system block diagram*

Equation 12-15 is solved for current divided by applied voltage as follows:

$$G(s)=\frac{I(s)}{V_a(s)}=\frac{1}{Ls+R+\frac{1}{SC}} \qquad (12\text{-}16)$$

which is then converted to root locus form:

$$G(s)=\frac{\frac{s}{L}}{s^2+\frac{R}{L}s+\frac{1}{LC}} \qquad (12\text{-}17)$$

The system transfer function gives insight to the system's performance by allowing key parameters to be calculated. The general form for a system transfer function is:

$$G(s)=\frac{\text{numerator}}{s^2+2\zeta\omega_n s+\omega_n^2} \qquad (12\text{-}18)$$

where ω_n is the natural frequency of the system, and ζ is known as the damping ratio.

The natural frequency of the system determines the speed of the system's response to a known input, and

the damping ratio determines how sluggish the system's response will be to a known input.

The two parameters of natural frequency and damping ratio can also be viewed from another perspective. Taking the denominator of the transfer function and setting it equal to zero is known as the "characteristic equation." If the characteristic equation is then solved for its roots, the location of these roots in the complex plane, which is a function of ω_n and ζ, will determine system performance, see figure 12-13.

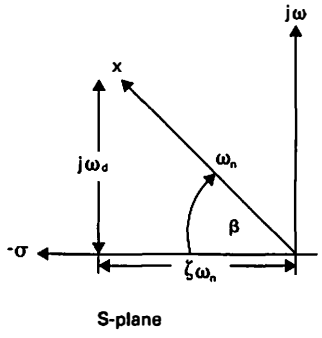


FIGURE 12-13 *The root of an under damped system plotted on the complex plane. This illustrates graphically the relationship between $\zeta\omega_n$, ω_n and ω_d.*

For equation 12-18, the characteristic equation is:

$$s^2+2\zeta\omega_n s+\omega_n^2=0 \qquad (12\text{-}19)$$

Using the quadratic formula to solve for the roots of the characteristic equation results in the following:

$$s=\frac{-2\zeta\omega_n \pm\sqrt{(2\zeta\omega_n)^2-4\omega_n^2}}{2} \qquad (12\text{-}20)$$

Which is equivalent to:

$$s=-\sigma \pm j\omega_d$$

Note: $\sigma = \zeta\omega_n$

Depending on the values of ω_n and ζ, six different types of values for the roots can result:

1. The roots are negative, real, and distinct.
2. The roots are negative, real, and repeated.
3. The roots are complex conjugate pairs with negative real parts.
4. The roots are purely imaginary.
5. The roots are complex conjugate pairs with positive real parts.
6. The roots have positive real parts.

The six distinct cases described will result in six different system responses, which are discussed in the next section.

System Response

The system response is defined as the output of a system with a known input. The response can be expressed in the Laplace domain, but is typically expressed in the time domain and graphed as a function of time.

Since a transfer function, $G(s)$, is defined as the ratio of the output, $X(s)$, divided by the input, $F(s)$, the output can be calculated by multiplying the input times the transfer function, as shown in equation 12-21.

$$X(s)=F(s)G(s)=F(s)\frac{X(s)}{F(s)} \qquad (12\text{-}21)$$

Hence, an expression for the input needs to be developed. The most common system input, and the one that will be used exclusively in this text, is the unit step input, symbolized by $u(t)$. A graph of the unit step function is shown in figure 12-14.

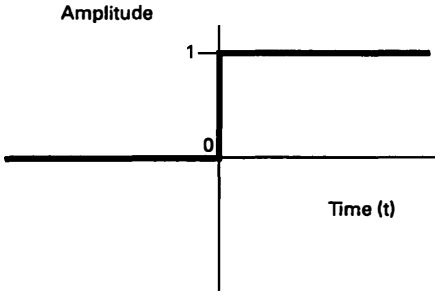


FIGURE 12-14 *Unit step function*

The step function is characterized by zero amplitude until time zero, and then a prompt jump to unity amplitude through time infinity. A physical example is turning on a light switch at time zero. The equation for the step function is as follows:

$$\begin{aligned} u(t)&=0, \quad t<0 \\ u(t)&=1, \quad t>0 \end{aligned} \qquad (12\text{-}22)$$

Also important is the Laplace transform for the unit step function, shown in equation 12-23.

$$L\{u(t)\}=\frac{1}{s} \qquad (12\text{-}23)$$

The system response in the Laplace domain for the three systems previously discussed can now be determined, assuming each system is excited by a step input. For the

translational system of figure 12-3, the system response to a step input is:

$$X(s) = F(s)G(s) = \frac{1}{s} \frac{\frac{1}{M}}{s^2 + \frac{B}{M}s + \frac{K}{M}} \qquad (12\text{-}24)$$

For the rotational system of figure 12-7, the system response to a step input is:

$$\theta(s) = T(s)G(s) = \frac{1}{s} \frac{\frac{1}{J}}{s^2 + \frac{B}{J}s + \frac{K}{J}} \qquad (12\text{-}25)$$

For the electrical system of figure 12-8, the system response to a step input is:

$$I(s) = V_a(s)G(s) = \frac{1}{s} \frac{\frac{s}{L}}{s^2 + \frac{R}{L}s + \frac{1}{LC}} \qquad (12\text{-}26)$$

Each of the above system responses can be converted to the time domain, and then plotted as a function of time.

Final Value Theorem

A useful calculation can be performed in the Laplace domain that will give the steady state value of the output (i.e., the system response) after the transients settle out. This steady state value is found using the final value theorem in the Laplace domain, and is analogous to taking the limit in the time domain as time approaches infinity.

Furthermore, the final steady state value calculated in the Laplace domain is the same value that would be calculated in the time domain. Given a function in the Laplace domain, $X(s)$, the steady state value, x_{ss}, is expressed as follows:

$$x_{ss} = \lim_{s \to 0} sX(s) \qquad (12\text{-}27)$$

For example, to find the steady state response for the translational system of figure 12-3, given a step input,

$$x_{ss} = \lim_{s \to 0} sX(s) = \lim_{s \to 0} sF(s)G(s)$$

$$\Rightarrow x_{ss} = \lim_{s \to 0} s \frac{1}{s} \frac{\frac{1}{M}}{s^2 + \frac{B}{M}s + \frac{K}{M}}$$

$$\Rightarrow x_{ss} = \frac{1}{M} \frac{M}{K} = \frac{1}{K}$$

System Time Responses

The previous section described system responses in the Laplace domain. Mathematical methods exist to convert a system response to the time domain and to produce a plot of the system response as a function of time.

The time response of the system indicates the steady state value that will be reached by the output, the path by which this steady state value will be achieved, and the time it will take to reach the steady state value.

Suppose it is desired to train a gun mount 2.5° to the right. The following illustrate the six possible responses that the gun mount can exhibit in response to the train order.

Overdamped Response

Figure 12-15 shows the response of an overdamped system. Note how the gun mount eventually reaches the desired bearing (2.5° to the right), but takes a relatively long time to get there.

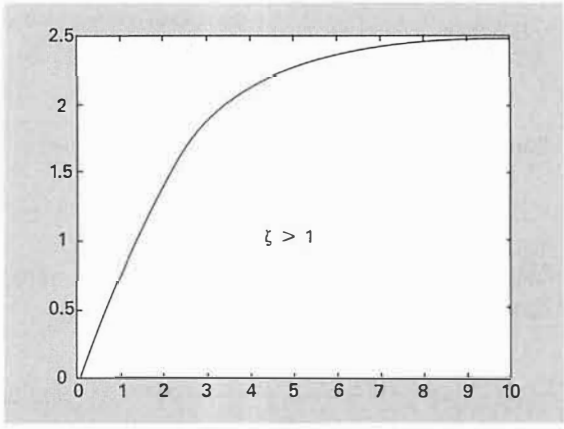


FIGURE 12-15 *Overdamped time response*

Overdamped systems are characterized by a damping ratio greater than 1 and roots of the characteristic equation that are negative, real, and distinct. Practically speaking, overdamped systems are seldom used due to the excessive time required for the system to respond. For a faster time response, less damping is used, as shown in the next two sections.

Critically Damped Response

Figure 12-16 shows the response for a critically damped system. In this case, the desired bearing is achieved much quicker as compared to the overdamped case. Critically damped systems are characterized by a damping ratio exactly equal to 1 and roots of the characteristic equation that are negative, real, and equal to each other.

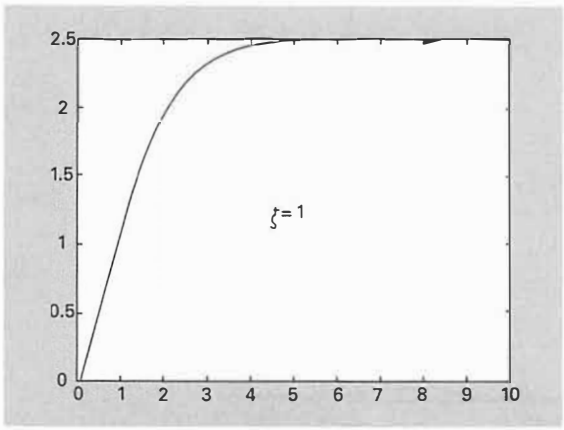


FIGURE 12-16 *Critically damped time response*

Underdamped Response

However, a quicker response may be desired than that given by the critically damped system. An underdamped response will, in fact, give a quicker response than the critically damped case, but will do so by initially passing through the desired value. This is called overshoot. Figure 12-17 shows the response for an underdamped system.

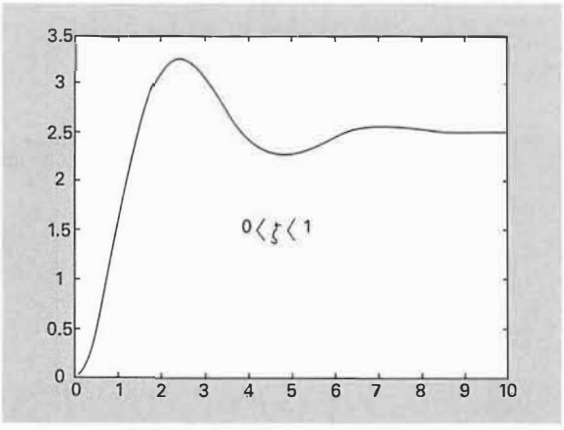


FIGURE 12-17 *Underdamped time response*

Recalling the roots of the characteristic equation, notice that with an underdamped system they take on special value. These will be useful in calculating many system characteristics. The roots of an underdamped system's characteristic equation are

$$s_{1,2} = -\zeta\omega_n \pm j\omega_d \qquad (12\text{-}28)$$

Notice that the real portion is negative and one-half the middle term of the characteristic equation. It is important to note the negative sign. For underdamped systems, $\zeta\omega_n$ and ζ are positive, *but the roots are not.* The imaginary portion of the roots is ω_d, the damped natural frequency. This is the system's actual frequency of oscillation. The only time that a system's actual natural frequency can be measured is in the undamped case, which will be discussed later.

Note the presence of the overshoot in Figure 12-17, where the response initially shoots past the desired value. This overshoot can be quantified and expressed as a percentage, called percent overshoot (%OS), that is defined as the percentage by which the final steady state value is exceeded on the initial peak of the response. This value can be measured from a plot of the time response, but if the value of the damping ratio, ζ, is known, the value of %OS can also be calculated as follows:

$$\begin{aligned} \%OS &= 100e^{\frac{-\pi\zeta}{\sqrt{1-\zeta^2}}} \\ &= 100e^{\frac{-4t_p}{t_s}} \\ &= 100\frac{X_{max} - X_{SS}}{X_{SS}} \end{aligned} \qquad (12\text{-}29)$$

An important parameter that is associated with %OS is the time the system takes to achieve the first and strongest peak oscillation. This is called peak time, t_p, and can also be measured from the time response or calculated using the following equation:

$$t_p = \frac{\pi}{\omega_n\sqrt{1-\zeta^2}} = \frac{\pi}{\omega_d} \qquad (12\text{-}30)$$

Also evident is the oscillation present prior to the system achieving the desired value. The time the system takes to actually settle out to the final steady state value can also be quantified and is known as settling time, t_s, which is more specifically defined as the period of time it takes a system to reach and maintain a value within 2% of the steady state value. The value of t_s can also be calculated when one knows the values of the damping ratio, ζ, and the natural frequency, ω_n, as follows:

$$t_s = \frac{4}{\zeta\omega_n} = \frac{4}{\sigma} \qquad (12\text{-}31)$$

Note: For under damped and critically damped systems $\sigma = \zeta\omega_n$. For over damped systems use σ which has the smaller absolute value.

Underdamped systems are characterized by damping ratios between 0 and 1, and roots of the characteristic equation that have a negative real part and imaginary parts that are complex conjugates of each other. It appears

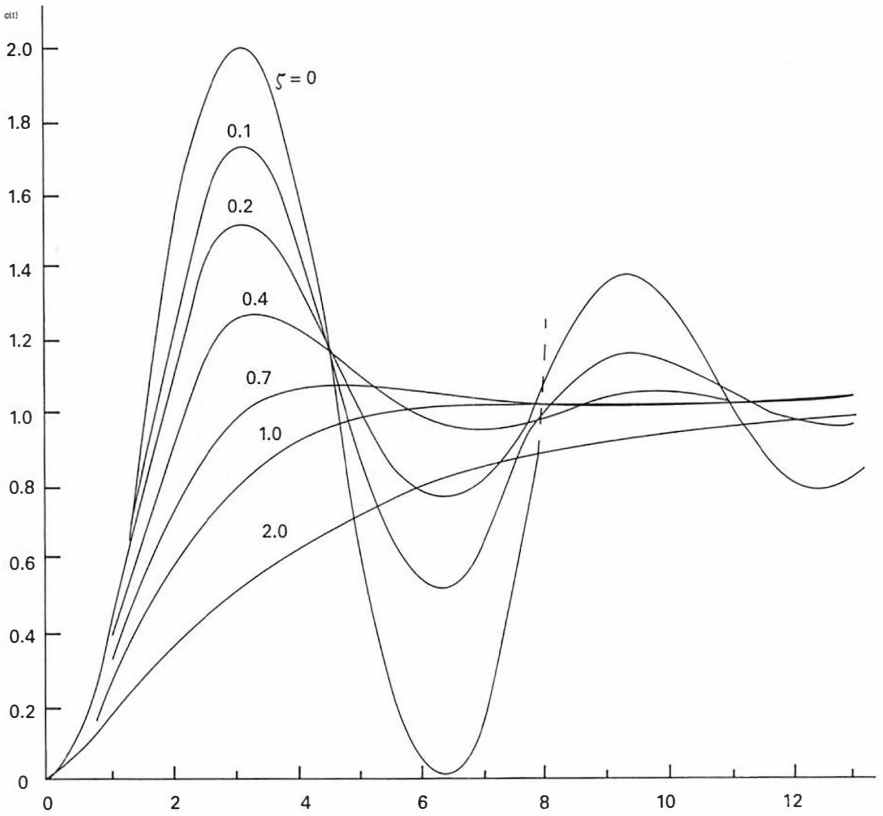


FIGURE 12-18 *Plot of system response and various damping ratios*

as though this type of response is undesirable due to the overshoot, but careful adjustment of damping and natural frequency value can minimize this overshoot and achieve a very quick response. The underdamped system is the most common type of system and the one that will be discussed in the greatest detail in this text.

Note: Equations for t_s, t_p, and %OS are approximations which apply to second order systems. The settling time equation is most accurate for under damped second order systems. A system's order is determined by the exponent, or highest power of s, in the transfer function's denominator. For example: $1/(s + 1)$ is a 1st order system, $1/(s^2 + 2s + 3)$ is a 2nd order system and $1/(s^3 + 4s + 2)$ is a 3rd order system. Higher order systems response can be approximated as 2nd order systems response and analyzed accordingly.

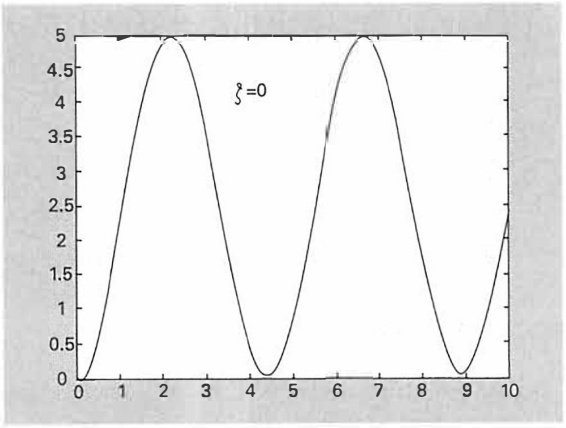


FIGURE 12-19 *Undamped time response*

Undamped Response

Figure 12-19 shows the response for an undamped system. In this case, the desired output is not achieved; rather, the response continues to oscillate about the desired value without ever achieving it. Undamped systems are characterized by a damping ratio equal to zero and roots of the characteristic equation that are imaginary only. These types of systems are not very common, and in actuality, express an ideal, since natural effects such as friction and air resistance will always be present to cause some type of damping.

Unstable Oscillating Response

Figure 12-20 shows the response for an unstable system. The system response not only fails to achieve the desired

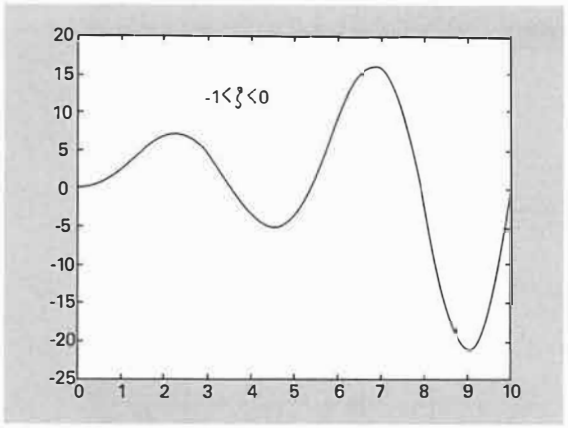


FIGURE 12-20 *Unstable oscillating response*

bearing but also oscillates wildly about the final value and quickly cycles out of control. An unstable system can cause injury and serious equipment damage. Unstable oscillating systems are characterized by a damping ratio less than 0, but greater than –1. The roots of the characteristic equation have complex conjugates with positive real parts.

Unstable Non-Oscillating Response

Figure 12-21 shows the response for an unstable system. The system response fails to achieve the desired bearing quickly, departs the final value, and is out of control. As with an unstable oscillating system, injury and serious equipment damage can occur. Unstable systems are characterized by a damping ratio less than negative one and roots with positive real parts.

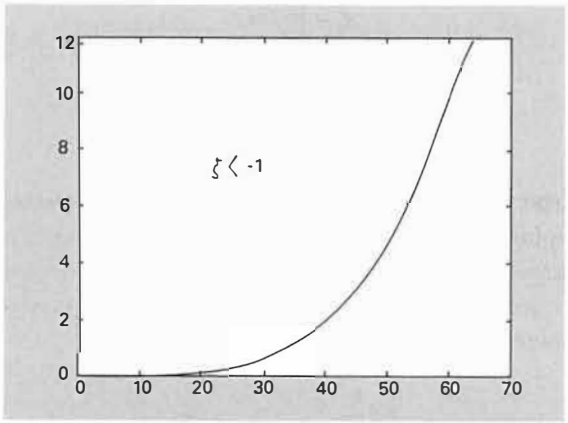


FIGURE 12-21 *Unstable non-oscillating response*

System Response Examples

EXAMPLE 12-1: Underdamped system. Recall the transfer function for the translational system of figure 12-3, given by equation 12-11:

$$G(s) = \frac{\frac{1}{M}}{s^2 + \frac{B}{M}s + \frac{K}{M}}$$

Assume this system represents a recoil system for a deck gun, with the following values for the mass, damping coefficient, and spring constant, respectively:

$$M = 2{,}000 \text{ kg}$$
$$B = 1{,}000 \text{ Ns/m}$$
$$K = 4{,}000 \text{ N/m}$$

Also, assume the input force for the system is

$$F = 10{,}000 \text{ N}$$

and is applied as a step input.

Knowing these parameters, it is desired to calculate the following:

1. The roots of the characteristic equation and determination of the system's time response
2. The system's final steady state value using the final value theorem
3. The values of the damping ratio, ζ, and the natural frequency, ω_n
4. The value of %OS
5. The value of t_p
6. The value of t_s

SOLUTION:

1. Substituting in for the transfer function gives:

$$G(s) = \frac{\frac{1}{2{,}000}}{s^2 + \frac{1{,}000}{2{,}000}s + \frac{4{,}000}{2{,}000}} = \frac{\frac{1}{2{,}000}}{s^2 + 0.5s + 2}$$

The roots of the characteristic equation are:

$$-0.25 \pm 1.39j$$

which indicates an underdamped system. The roots also indicate that $\zeta\omega_n$ equals 0.25 and ω_d equals 1.39.

2. The steady state value is calculated using the final value theorem of equation 12-27:

$$x_{ss} = \lim_{s \to 0} sX(s) = \lim_{s \to 0} sR(s)G(s)$$

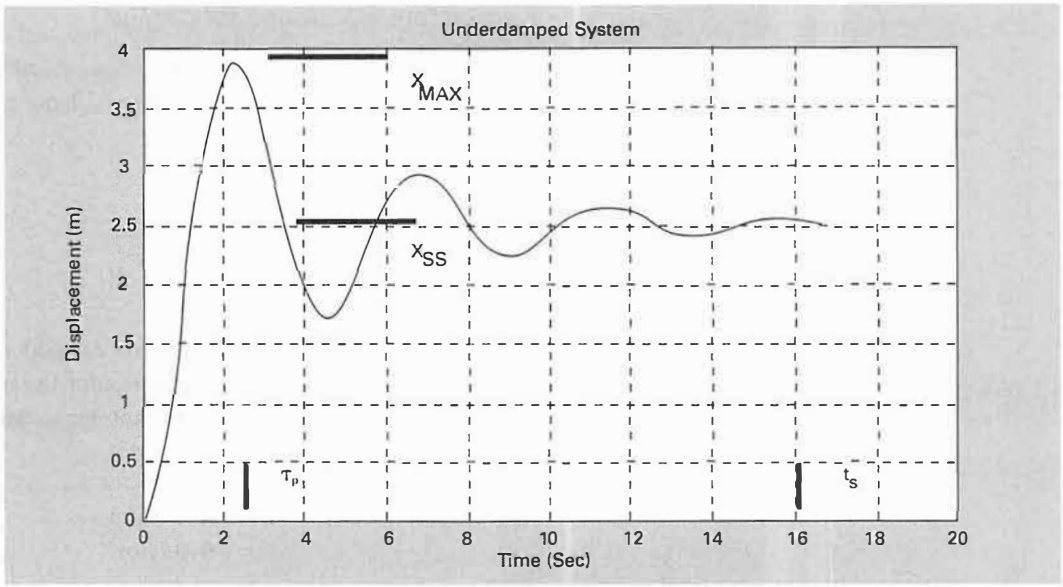


FIGURE 12-22 *Example 12-1 time response*

Because the input is a step input,

$$R(s) = \frac{F}{s} = \frac{10{,}000}{s}$$

$$x_{ss} = \lim_{s\to 0} sF(s)\frac{\frac{1}{M}}{s^2 + \frac{B}{M}s + \frac{K}{M}}$$

$$\Rightarrow x_{ss} = \lim_{s\to 0} s\frac{10{,}000}{s}\frac{\frac{1}{2{,}000}}{s^2 + 0.5s + 2}$$

$$= \frac{5}{(0)^2 + 0.5(0) + 2} = 2.5 \text{ m}$$

3. ζ and ω_n are found from the system transfer function:

$$\omega_n^2 = 2 \Rightarrow \omega_n = 1.41$$

$$2\zeta\omega_n = 0.5 \Rightarrow \zeta\omega_n = 0.25 \Rightarrow \zeta = 0.1768$$

Since ζ is between 0 and 1, the underdamped time response is confirmed.

4. %OS is calculated as follows:

$$\%OS = 100e^{\frac{-\pi(0.1768)}{\sqrt{1-(0.1768)^2}}} = 56.8\%$$

5. t_p is calculated as follows:

$$t_p = \frac{\pi}{\omega_d} = \frac{\pi}{1.39} = 2.26 \text{ sec}$$

6. t_s is calculated as follows:

$$t_s = \frac{4}{\sigma} = \frac{4}{\zeta\omega_n} = \frac{4}{0.25} = 16.0 \text{ sec}$$

In other words, this system would take 16 seconds to settle to its steady state value of 2.5 meters.

A plot of the time response is shown in figure 12-22. Note how the predicted parameters and system characteristics from the equations above are illustrated by the actual response of the system.

Note that t_s, t_p, x_{ss}, and %OS can all be ascertained from the graph of the output.

$$t_p = 2.25 \text{ sec}$$
$$t_s = 16 \text{ sec}$$
$$x_{ss} = 2.5$$
$$\%OS = 100\frac{3.9 - 2.5}{2.5} = 56\%$$

EXAMPLE 12-2: Underdamped system with increased damping coefficient. Given the same system parameters as in Example 12-1, but this time with the damping coefficient increased from 1,000 Ns/m to 4,000 Ns/m, it is desired to examine the effect on the output parameters:

$$G(s) = \frac{\frac{1}{2{,}000}}{s^2 + \frac{4{,}000}{2{,}000}s + \frac{4{,}000}{2{,}000}} = \frac{\frac{1}{2{,}000}}{s^2 + 2s + 2}$$

SOLUTION:

1. roots = $-1 \pm j$, which again indicates an underdamped system. The roots also indicate $\zeta\omega_n = 1$ and $\omega_d = 1$.
2. The steady state value is calculated using the final value theorem:

$$x_{ss} = \lim_{s\to 0} s \frac{10{,}000}{s} \frac{\frac{1}{2{,}000}}{s^2+2s+2}$$

$$\Rightarrow x_{ss} = \frac{5}{(0)^2+0.5(0)+2} = 2.5 \text{ m}$$

It is observed that the steady state value does not change from the previous example.

3. ζ and ω_n are found from the system transfer function:

$$\omega_n^2 = 2 \Rightarrow \omega_n = 1.41$$

$$2\zeta\omega_n = 2 \Rightarrow \zeta\omega_n = 1 \Rightarrow \zeta = 0.707$$

4. %OS is calculated as follows:

$$\%OS = 100e^{\frac{-\pi(0.707)}{\sqrt{1-(0.707)^2}}} = 4.33\%$$

5. t_p is calculated as follows:

$$t_p = \frac{\pi}{\omega_d} = \frac{\pi}{1} = 3.14 \text{ sec}$$

6. t_s is calculated as follows:

$$t_s = \frac{4}{\sigma} = \frac{4}{\zeta\omega_n} = \frac{4}{1} = 4.0 \text{ sec}$$

The above parameters verify the time domain response shown in figure 12-23.

EXAMPLE 12-3: Underdamped system with increased spring constant. Now, if the damping coefficient was changed back to 1,000 Ns/m, but the spring constant, K, was increased to 5,000 N/m, the following changes would be observed:

$$G(s) = \frac{\frac{1}{2{,}000}}{s^2 + \frac{1{,}000}{2{,}000}s + \frac{5{,}000}{2{,}000}} = \frac{\frac{1}{2{,}000}}{s^2+0.5s+2.5}$$

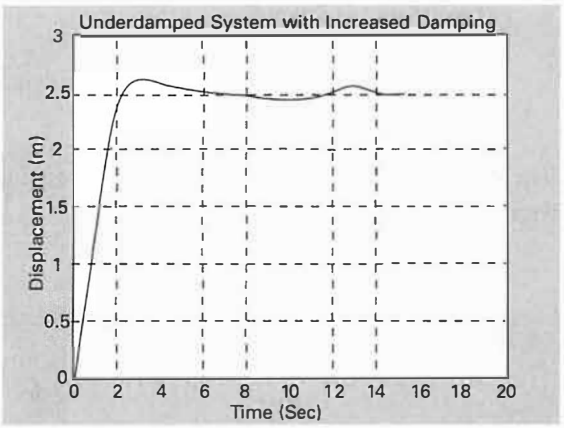


FIGURE 12-23 *Example 12-2 time response*

SOLUTION:

1. roots = $-0.25 \pm 1.56j$, which again indicates an underdamped system. The roots also indicate $\zeta\omega_n = 0.25$ and $\omega_d = 1.56$.
2. The steady state value is calculated using the final value theorem:

$$x_{ss} = \lim_{s\to 0} s \frac{10{,}000}{s} \frac{\frac{1}{2{,}000}}{s^2+0.5s+2.5}$$

$$\Rightarrow x_{ss} = \frac{5}{(0)^2+0.5(0)+2.5} = 2 \text{ m}$$

and it is observed that the steady state value does in fact change from the previous two examples.

3. ζ and ω_n are found from the system transfer function:

$$\omega_n^2 = 2.5 \Rightarrow \omega_n = 1.58$$

$$2\zeta\omega_n = 0.5 \Rightarrow \zeta\omega_n = 0.25 \Rightarrow \zeta = 0.158$$

Since ζ is between 0 and 1, the presence of an underdamped system is again confirmed.

4. %OS is calculated as follows:

$$\%OS = 100e^{\frac{-\pi(0.158)}{\sqrt{1-(0.158)^2}}} = 60.47\%$$

5. t_p is calculated as follows:

$$t_p = \frac{\pi}{\omega_d} = \frac{\pi}{1.56} = 2.01 \text{ sec}$$

6. t_s is calculated as follows:

$$t_s = \frac{4}{\sigma} = \frac{4}{\zeta\omega_n} = \frac{4}{0.25} = 16.0 \text{ sec}$$

The above parameters verify the time domain response shown in figure 12-24.

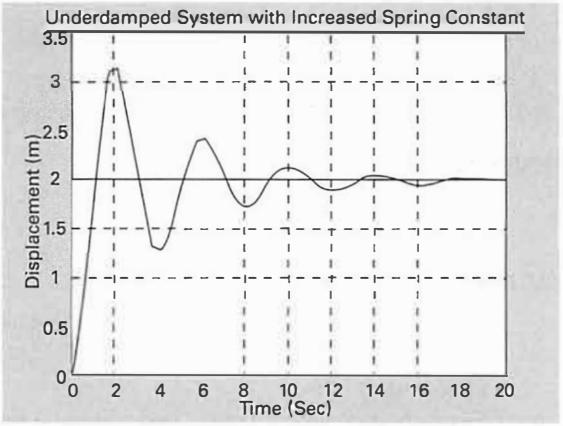


FIGURE 12-24 *Example 12-3 time response*

The Effect of System Components

An interesting relationship between springs, dampers, and time response can be found. For instance, what effect will putting stiffer springs on a car have? Returning to our generic transfer function for a translational system (mass spring damper), in equation 12-11, we can algebraically break it into more useful forms. Students should note that similar results could be obtained through analysis of the rotational and electrical systems.

$$G(s) = \frac{\frac{1}{M}}{s^2 + \frac{B}{M}s + \frac{K}{M}}$$

Remember that the denominator of the transfer function is the characteristic equation, $s^2 + 2\zeta\omega_n s + \omega_n^2$, and that the roots of the characteristic equation are $-\zeta\omega_n \pm j\omega_d$.

Relating the equations, there are two immediate observations:

$$\zeta\omega_n = \frac{B}{2M}$$

and

$$\omega_n^2 = \frac{K}{M} \quad \text{or} \quad \omega_n = \sqrt{\frac{K}{M}}$$

Of interest when discussing system response is damping ratio, which relates directly to overshoot and settling time.

$$\zeta = \frac{\zeta\omega_n}{\omega_n} = \frac{\frac{B}{2M}}{\sqrt{\frac{K}{M}}} = \frac{B}{2\sqrt{MK}}$$

$$t_s = \frac{4}{\zeta\omega_n} = \frac{4}{\frac{B}{2M}} = \frac{8M}{B}$$

A third time response characteristic, time to peak, can be calculated by finding the damped natural frequency and then inserting that into the equation for time to peak.

$$\omega_n^2 = \omega_d^2 + (\zeta\omega_n)^2$$

$$\omega_d^2 = \omega_n^2 - (\zeta\omega_n)^2 = \frac{K}{M} - \left(\frac{B}{2M}\right)^2$$

$$\omega_d = \sqrt{\frac{K}{M} - \left(\frac{B}{2M}\right)^2}$$

$$t_p = \frac{\pi}{\omega_d} = \frac{\pi}{\sqrt{\frac{K}{M} - \left(\frac{B}{2M}\right)^2}}$$

Finally, the steady state output, x_{ss}, can be calculated using the final value theorem. Note that as before, we have assumed a step input.

$$x_{ss} = s\frac{F}{s}\frac{\frac{1}{M}}{s^2 + \frac{B}{M}s + \frac{K}{M}} = \frac{F}{K}$$

What does this tell us about the time response of a system?

Table 12-2 gives some general characteristics with respect to *M*, *B*, and *K*. Because the force, *F*, affects only the steady state output and not the other characteristics, it is not included in the table.

TABLE 12-2 *Affect of spring constant and damping constant on time response while other parameters held constant*

	$\zeta\omega_n$	ω_n	ω_d	ζ	%OS	t_p	t_s	x_{ss}
K↑	↔	↑	↑	↓	↑	↓	↔	↓
K↓	↔	↓	↓	↑	↓	↑	↔	↑
B↑	↑	↔	↓	↑	↓	↑	↓	↔
B↓	↓	↔	↑	↓	↑	↓	↑	↔
M↑	↔	↓	↓	↓	↑	↑	↔	↔
M↓	↔	↑	↑	↑	↓	↓	↔	↔

Imagine a car's suspension after changing to stiffer springs. This will increase the percent overshoot (the actual displacement may decrease) and decrease the time to peak, resulting in a harsher ride. To counteract this and soften the ride, you might change to stiffer shocks (damper); this would reduce the overshoot and increase time to peak.

Selecting the right combination is critical to producing the desired characteristics. If we describe time to peak and percent overshoot as comfort characteristics and settling time as a rate characteristic, there are some general statements that can be made.

1. Changing the spring will not affect the rate of the response, but it will affect comfort.
2. Changing the damper will affect both the rate of response and comfort.

In many cases it is not realistic to alter the time response characteristics of a system by changing the components. This is unnecessarily complicated and expensive. The alternative is to add feedback. Feedback allows us to effectively modify the time response of a system without changing the components. Furthermore, feedback will allow us to adjust the time response of a system for a wide range of operating conditions.

Feedback

Feedback is defined as the comparison of the system response or output with the system input (also known as the desired output). Without this comparison, the system operator can never be sure that the system is actually achieving the desired output. An example of a system block diagram with unity or negative feedback is shown below in figure 12-25. Note how the system input $R(s)$ is fed through a device called a summing junction and is then fed forward to the usual system transfer function, $G(s)$ to the output, $C(s)$. However, the output is sensed by a sensing device and then fed back to the summing junction.

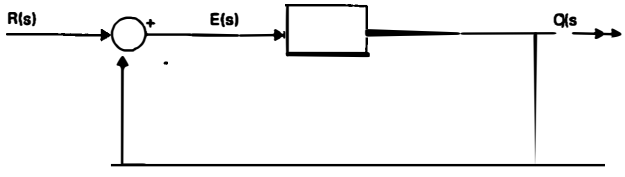


FIGURE 12-25 *System block diagram with negative feedback*

The comparison of output and input takes place in the summing junction. For the negative feedback case, the system input is given a plus sign, and the system output is given a minus sign, resulting in the subtraction of the output from the input. The difference of these two quantities is called the error, $E(s)$. When the error is zero, the system acts to maintain the status quo. When the error is non-zero, the system changes to bring the error to zero. For the positive feedback case, the system input is given a plus sign, and the system output is also given a plus sign. The positive feedback case is less common and results in the addition of the input and the output. This text uses the negative feedback case for the remainder of this chapter.

Most systems use feedback, and much of the feedback and comparison processes can be done through use of software as part of the control system. However, one may find systems without feedback, called open-loop systems. These types of systems are discussed below.

Open-Loop Systems

An open-loop system is one without feedback, as shown in figure 12-26. In other words, an input is provided to the system, but no check on the system's output or comparison with the input is performed by the system itself. It is assumed that the system is performing as required, or the system performance is maintained and checked by an operator. In general, open-loop systems are simpler, easier to design, and less expensive to build.

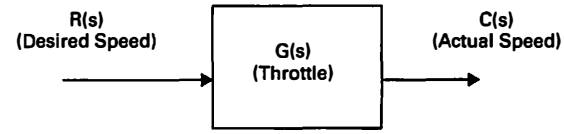


FIGURE 12-26 *Open-loop vehicle speed control block diagram*

However, there is no mechanism to automatically correct the system output should the expected output not be achieved due to some unforeseen circumstance. If we wish to change the output of the system without changing the input, we have to add some amplification. This is commonly described as forward-path amplification, and we will use the symbol K_{amp} to describe it.

An example of an open-loop system is a cruise control system, where the throttle moves to a predetermined position for a given speed.

This system may perform quite well on a flat road during a calm day. However, when the vehicle encounters an unexpected event called a disturbance, such as the hill shown in figure 12-27, the throttle position remains the same, and speed drops.

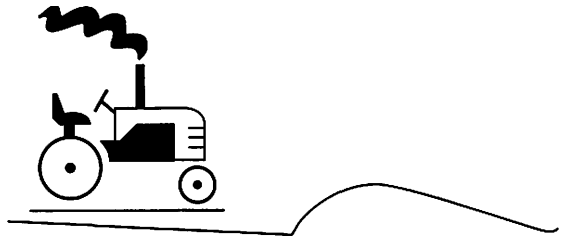


FIGURE 12-27 *Open-loop system encountering a disturbance*

One might argue that once the hill is encountered, the driver of the tractor should deactivate the cruise control and manually increase the throttle position to account for the hill. In actuality, this is feedback and closed-loop control provided by the driver. In other words, the driver performs the functions of feedback and comparison of the output with the desired input. Although open-loop systems are used in many applications, human beings are accustomed to periodically monitoring their performance and, if necessary, making adjustments to the input. This is one of the many reasons that data logs are taken on ships, aircraft, and submarines at sea.

Block Diagram

Prior to examining specific transfer functions for feedback lopes, a discussion on the utility of block diagrams is needed. When building a mathematical model of a system it is useful to draw a diagram showing the relationship between the different components. This is accomplished with block diagrams that help break down the problem into manageable components.

Consider a simple spring. Hook's Law for a spring is $F = K \cdot X$. Given a force applied to the spring, the spring's displacement can be found; or given an applied displacement the resulting force can be obtained. Figure 12-28 displays this as a block diagram.

This spring example illustrates that the fundamental operation of the block diagram is multiplication. The output of the block is the product of the input and the block itself. Figure 12-29 shows a pair of blocks connected and the resulting output.

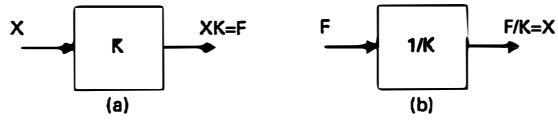


FIGURE 12-28 *The block diagram for a spring showing displacement as the input and force as the output (a) and force as the input and displacement as the output (b).*

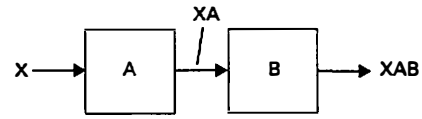


FIGURE 12-29 *The output is the product of the input and the block.*

Addition and Subtraction

Addition (and subtraction) is handled by a summing junction on the outputs of the blocks.

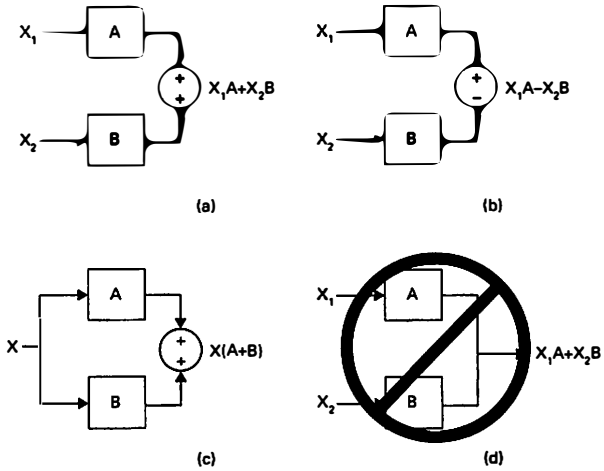


FIGURE 12-30 *Three correct and one incorrect examples of block diagram addition. Figure (d) is incorrect because there is no summing junction connecting the outputs of blocks A and B so one does not know how to combine them.*

Figure 12-30b shows that subtraction is accomplished by changing one or more of the signs on the summing junctions. Figure 12-30d shows that outputs cannot be combined without a summing junction. Note: summing junctions are not limited to just two inputs.

Block Simplification

Block diagrams can be simplified. Multiple blocks and summing junctions can be reduced to a single block which is equivalent to the initial un-simplified diagram.

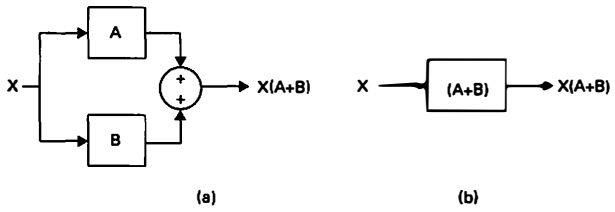


FIGURE 12-31 *The diagram in (a) can be simplified to the diagram in (b) and they are equivalent.*

Open-Loop Transfer Function (OLTF) Example

EXAMPLE 12-4: System with forward-path amplification. Given the system in figure 12-32 and an input of $5u(t)$, find the steady state output.

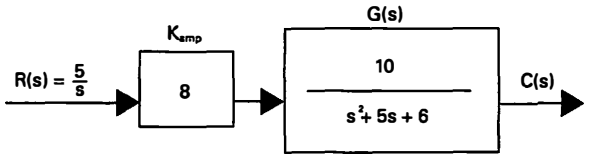


FIGURE 12-32 *Example 12-4 open loop with forward-path amplification*

SOLUTION: A new transfer function can be obtained by simply multiplying the blocks. This becomes

$$G_{OLTF} = K_{amp}G(s) = \frac{80}{s^2+5s+6}$$

To find the steady state output we apply the final value theorem:

$$\begin{aligned} x_{ss} &= \lim_{s\to 0} sR(s)G_{OLTF} \\ &= \lim_{s\to 0} s\frac{5}{s}\frac{80}{s^2+5s+6} \\ &= \frac{400}{6} = 66.7 \end{aligned}$$

Closed-Loop Systems

A closed-loop system is one with feedback and a comparison between the input (or desired output) and the actual output. As systems become more and more complicated and capable, and as there is an effort to reduce the number of personnel on ships and submarines, it is unrealistic to expect human beings to continually monitor systems as they have traditionally.

Many modern systems are closed-loop systems complete with feedback, calculation of the system error and adjustment of system performance in response to this error. All of these functions can be performed via software embedded in the system's microprocessor, thereby saving space and complexity in the system's design.

A transfer function can be derived for the complete closed-loop system. Figure 12-33 shows a generic closed-loop system, with a transfer function, $H(s)$, in the feedback path. This may be a simple rational number, called a gain, or a more complex polynomial.

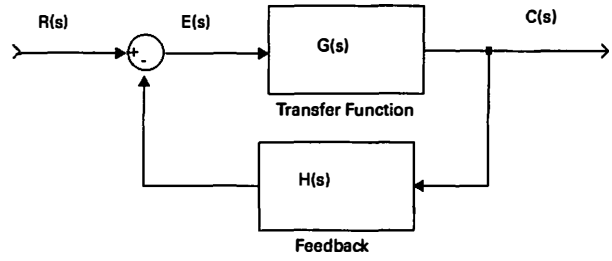


FIGURE 12-33 *Closed-loop system with negative feedback transfer function, H(s)*

Figure 12-33 specifically illustrates a negative feedback loop. The "feedback loop" refers to taking a portion of the output signal, $C(s)$, and combining it with the input signal $R(s)$, via the $H(s)$ block and the summing junction. "Negative" refers to the negative sign in the summing junction. A "positive feedback loop" is illustrated in figure 12-34. Note the sign difference in the summing junction.

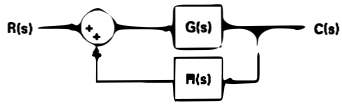
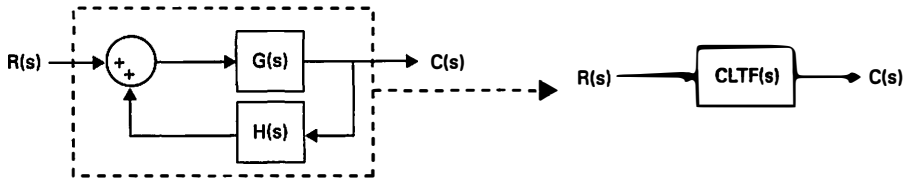


FIGURE 12-34 *A positive feedback loop*

The block diagram in figures 12-33 and 12-34 are fundamentally different from the erroneous block diagram depicted in figure 12-30d because of the direction of the arrow on the output side.

Positive and negative feedback loops can be reduced to a single block and function. This function is the "Closed Loop Transfer Function" and for the negative case can be derived as follows:

$$C(s) = E(s)G(s) \qquad (12\text{-}32)$$



$$CLTF_{negative}(s) = \frac{G(s)}{1 + G(s)H(s)}$$

FIGURE 12-35 *A negative feedback loop reduced to a single block and function*

This can also be expressed for $E(s)$:

$$E(s) = R(s) - H(s)C(s) \qquad (12\text{-}33)$$

Substituting equation 12-33 into 12-32 gives:

$$C(s) = [R(s) - H(s)C(s)]G(s) \qquad (12\text{-}34)$$

Expanding and rearranging gives:

$$C(s) = R(s)G(s) - H(s)C(s)G(s)$$

$$C(s)[1 + G(s)H(s)] = R(s)G(s)$$

$$\frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)}$$

Therefore, the expression for the closed-loop transfer function (CLTF) is:

$$\text{CLTF} = \frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)} \qquad (12\text{-}35)$$

Closed-Loop Transfer Function (CLTF) Examples

EXAMPLE 12-5: System with unity feedback. Given the system shown in figure 12-36, find the CLTF.

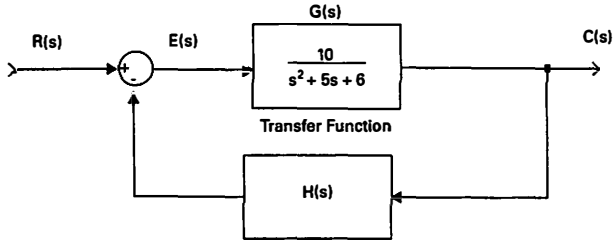


FIGURE 12-36 *Example 12-5 system with unity feedback*

SOLUTION: Since this is a case of unity feedback, $H(s) = 1$ and equation 12-35 gives:

$$\text{CLTF} = \frac{\frac{10}{s^2+5s+6}}{1 + \frac{10}{s^2+5s+6}}$$

Multiplying top and bottom by the polynomial $s^2 + 5s + 6$ gives:

$$\text{CLTF} = \frac{10}{s^2+5s+6+10}$$

This simplifies to:

$$\text{CLTF} = \frac{10}{s^2+5s+16}$$

EXAMPLE 12-6: System with forward-path amplification and unity feedback. Given the system in figure 12-37, find the CLTF.

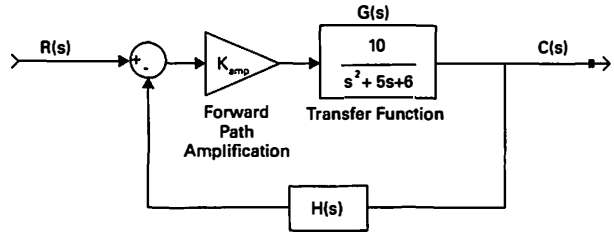


FIGURE 12-37 *Forward path-amplification and unity feedback*

SOLUTION: In this case, equation 12-35 gives:

$$\text{CLTF} = \frac{\frac{10K_{amp}}{s^2+5s+6}}{1 + \left(\frac{10K_{amp}}{s^2+5s+6}\right)}$$

Multiplying top and bottom by the polynomial $s^2 + 5s + 6$ gives:

$$\text{CLTF} = \frac{10K_{amp}}{s^2+5s+6+10K_{amp}}$$

Assuming K_{amp} has a value of 2, the following CLTF results:

$$\text{CLTF} = \frac{20}{s^2+5s+26}$$

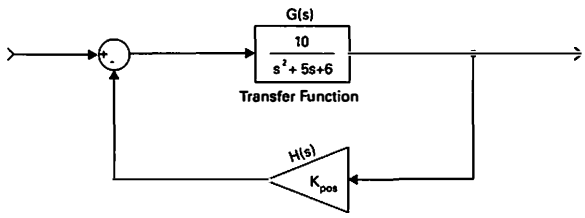


FIGURE 12-38 *Example 12-7 system with position feedback*

EXAMPLE 12-7: System with position feedback. Given the system in figure 12-38, find the CLTF.

SOLUTION: In this case, it is assumed that the input is a desired position and that the output is the actual position produced by the system. The feedback samples the output position value and multiplies it by a rational number or gain, K_{pos}. This feedback gain can be viewed as a number that scales the output so that the proper comparison between output and input is made at the summing junction.

In this case, equation 12-35 gives:

$$CLTF = \frac{\frac{10}{s^2+5s+6}}{1+K_{pos}\left(\frac{10}{s^2+5s+6}\right)}$$

Multiplying top and bottom by the polynomial $s^2 + 5s + 6$ gives:

$$CLTF = \frac{10}{s^2+5s+6+10K_{pos}}$$

Assuming K_{pos} has a value of 0.5, the following CLTF results:

$$CLTF = \frac{10}{s^2+5s+11}$$

EXAMPLE 12-8: System with rate and position feedback. Frequently, it is also desired to have information on how quickly a system is changing its position in addition to real-time information on where the current position of the system actually is. Even if position is the only output parameter being sensed by the system, the system speed is easily calculated by taking the derivative of position. Since this new parameter is now a rate of change of position, and since it is in the feedback path of the system, it is called rate feedback. Recall that in the Laplace domain, the derivative is taken by the multiplication by s.

Figure 12-39 shows a system with rate and position feedback. Note that the output, $C(s)$, is sensed and fed back through a transfer function containing the position feedback gain, K_{pos}, and the rate feedback, K_{rate}[2].

Given the system shown in Figure 12-39, find the CLTF.

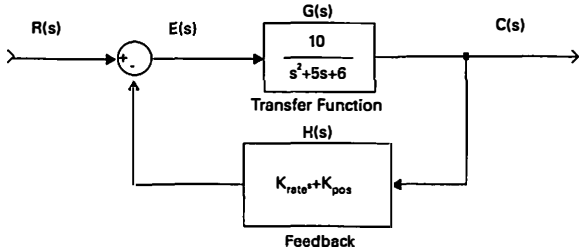


FIGURE 12-39 *Example 12-8 system with rate and position feedback*

SOLUTION: In this case, equation 12-35 gives:

$$CLTF = \frac{\frac{10}{s^2+5s+6}}{1+(K_{rate}s+K_{pos})\left(\frac{10}{s^2+5s+6}\right)}$$

Multiplying top and bottom by the polynomial $s^2 + 5s + 6$ gives:

$$CLTF = \frac{10}{s^2+(5+10K_{rate})s+6+10K_{pos}}$$

Assuming K_{pos} has a value of 0.5, and K_{rate} has a value of 0.8, the following CLTF results:

$$CLTF = \frac{10}{s^2+13s+11}$$

Error

In the previous section on closed-loop systems, the error signal, $E(s)$, was defined as the subtraction of the output signal from the input signal. When an input signal is first applied, $E(s)$ will increase, but the system will respond to drive the error to a minimal value after a period of time. Ideally, this is zero, but in some cases, the error reaches a steady state value other than zero. If a system produces a final steady state error signal other than zero, the system is said to have steady state error, symbolized by e_{ss}.

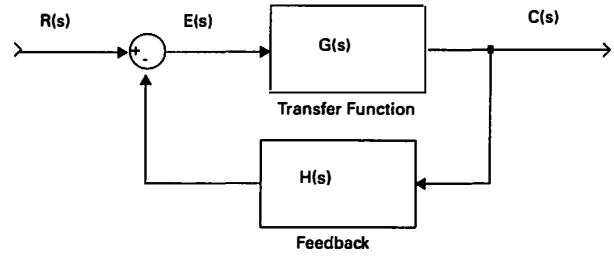


FIGURE 12-40 *System block diagram with negative feedback*

Given the system shown in figure 12-40 with a known input $R(s)$, it is desired to find an expression for the error signal, $E(s)$.

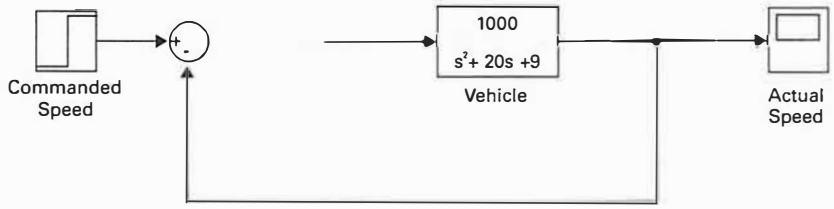


FIGURE 12-41 *Example 12-9 vehicle speed control system*

Recall equations 12-32 and 12-33:

$$C(s)=E(s)G(s)$$
$$E(s)=R(s)-H(s)C(s)$$

Substituting the expression for $C(s)$ from equation 12-32 into equation 12-33 gives:

$$E(s)=R(s)-H(s)E(s)G(s) \qquad (12\text{-}36)$$

Rearranging and solving for $E(s)$ gives:

$$E(s)+H(s)E(s)G(s)=R(s)$$
$$E(s)[1+G(s)H(s)]=R(s)$$
$$E(s)=\frac{R(s)}{1+G(s)H(s)} \qquad (12\text{-}37)$$

So knowing the input signal $R(s)$ and the forward path and feedback transfer functions $G(s)$ and $H(s)$, an expression for the error signal $E(s)$ can be found. Furthermore, a value can be found for this signal in steady state by using the final value theorem:

$$e_{ss}=\lim_{s\to 0} sE(s) \qquad (12\text{-}38)$$

Substituting in for $E(s)$ gives:

$$e_{ss}=\lim_{s\to 0} s\frac{R(s)}{1+G(s)H(s)} \qquad (12\text{-}39)$$

EXAMPLE 12-9: Steady state error for a speed control system. Given the speed control system shown in figure 12-41, with an input speed of 55 mph given as a step input, calculate the steady state error for the system.

SOLUTION: Referring to final value theorem, the steady state error is calculated by:

$$e_{ss}=\lim_{s\to 0} s\frac{R(s)}{1+G(s)H(s)}$$

In this case, $R(s) = 55/s$, since the input is given as a step input. $H(s) = 1$ since the system is a unity feedback system.

Substituting in gives:

$$e_{ss}=\lim_{s\to 0} s\frac{\frac{55}{s}}{\left(1+\frac{1{,}000}{s^2+20s+9}\right)}$$

$$e_{ss}=\frac{55}{\left(1+\frac{1{,}000}{0+0+9}\right)}=\frac{55}{112.1}=0.49 \text{ mph}$$

So, although the desired speed is 55 mph, the nature of the system dictates that there exists a steady state error of about 0.49 mph. Figure 12-42 is a plot of the vehicle speed. Notice that the steady state speed is slightly less than 55 mph.

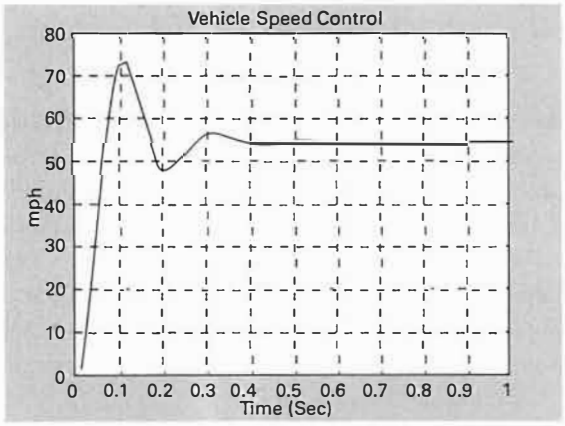


FIGURE 12-42 *Example 12-9 plot of vehicle speed*

Figure 12-43 is a plot of the error signal, $E(s)$. Note that upon application of the step input, the error signal is large, but the system responds by reducing the error to near zero. At steady state, however, $E(s)$, never quite reaches zero, but steadies out at 0.49 mph.

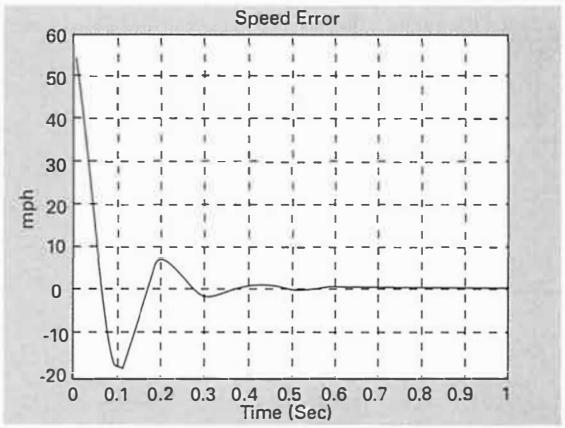


FIGURE 12-43 *Example 12-9 error time response*

System Controllers or Compensators

Having reviewed many aspects of system performance such as speed of response, percent overshoot, and steady state error, it is desired to develop a method to improve these performance parameters. As mentioned in the section on feedback, one method is to adjust the rate and position feedback gains.

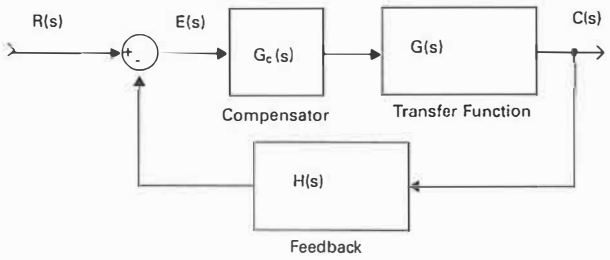


FIGURE 12-44 *System block diagram with controller,* $G_c(s)$

A more common method, however, is to design a controller or compensator in the forward-path loop to improve the system performance. Figure 12-44 shows a typical closed-loop system with a controller installed. Note how the controller, $G_c(s)$, takes the error signal from the summing junction $E(s)$, modifies it, and sends it to the plant or system transfer function, $G(s)$. By adjusting $G_c(s)$, system performance parameters can be improved.

This section will discuss four types of controllers:

- Proportional control
- Proportional-derivative (PD) control
- Proportional-integral (PI) control
- Proportional-integral-derivative (PID) control

Proportional Control

Figure 12-45 shows a block diagram of a system with proportional control. This is the simplest means of providing control to a system and consists merely of placing a gain in the forward path by which $E(s)$ is multiplied prior to being sent to the system transfer function, $G(s)$. This gain is a rational number, K_{amp}, that may be greater than or less than 1, but is always greater than 0.

Note how the gain or proportional control can be represented by a right-facing triangle symbol. The use of proportional or gain control can improve the system response, but it is limited in its ability to do so. It can also affect stability, either positively or negatively.

Proportional-Derivative (PD) Control

Figure 12-46 shows a block diagram of the cruise control system of figure 12-41 with PD control added to the system. As is indicated in the block diagram, PD control has a proportional gain, in this case, A, but it also has a derivative function with an associated gain, B. By taking the derivative of the error signal $E(s)$, the system can better anticipate the changes in $E(s)$, and respond quicker.

Thus, PD control is used to improve system response—that is, reduce time to peak, t_p, and settling time, t_s. Adjustment of the gain values A and B allows the designer to finely tune the system's performance.

Proportional-Integral (PI) Control

Figure 12-47 shows a block diagram of the cruise control system of figure 12-41 with PI control added to the system. As is indicated in the block diagram, PI control has a proportional gain, in this case, A, but it also has an integral function with an associated gain, C. By taking the integral of the error signal $E(s)$, the system has an improved capability to steady out at the desired steady state value.

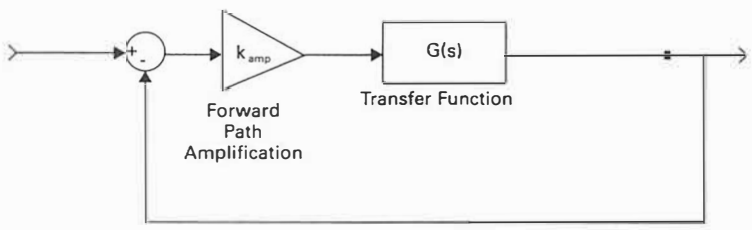


FIGURE 12-45 *System block diagram with proportional control*

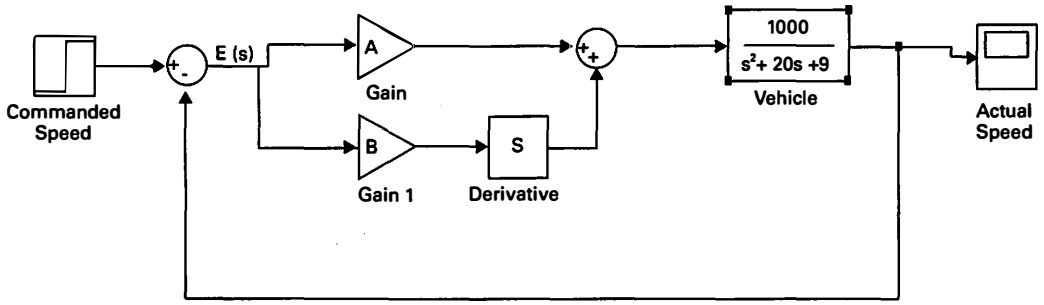


FIGURE 12-46 *System block diagram with PD control*

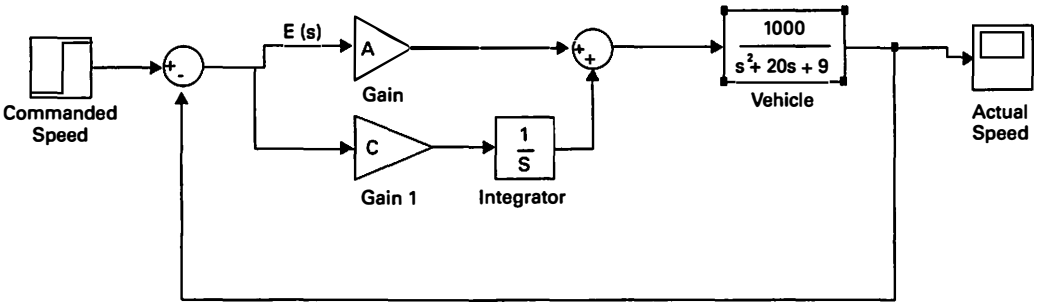


FIGURE 12-47 *System block diagram with PI control*

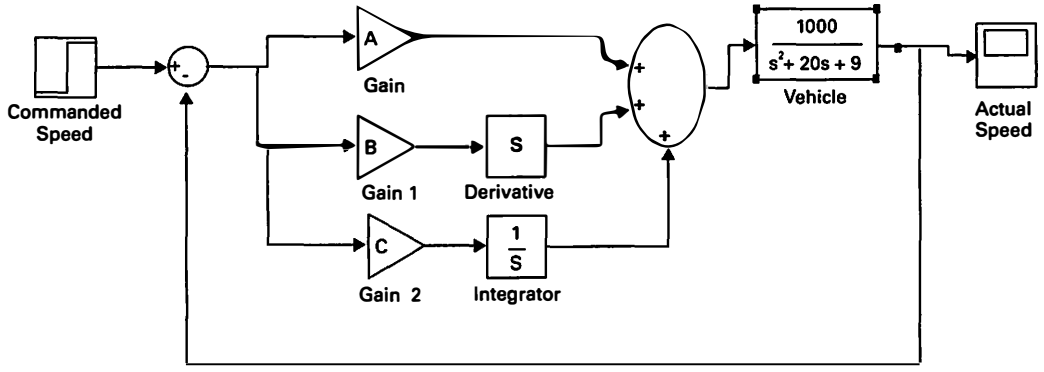


FIGURE 12-48 *System block diagram with PID control*

Thus, PI control is used to reduce or eliminate steady state error. However, PI control does nothing to improve system performance, and may in fact degrade it.

The steady state error, e_{ss}, of the system can be calculated using equation 12-39. In this case, the entire forward-path gain consisting of the product of the plant ("Vehicle") and the PI controller are substituted for $G(s)$:

$$e_{ss} = \lim_{s\to 0} s \frac{\frac{55}{s}\frac{1}{1+\left(A+\frac{C}{s}\right)\left(\frac{1{,}000}{s^2+20s+9}\right)}}$$

$$e_{ss} = \lim_{s\to 0} \frac{55s}{s+(As+C)\left(\frac{1{,}000}{s^2+20s+9}\right)} = \frac{0}{C\left(\frac{1{,}000}{9}\right)}$$

$$e_{ss} = 0.0 \text{ mph}$$

As can be seen from the above example, the steady state error has been reduced to zero by use of a PI controller.

Proportional-Integral-Derivative (PID) Control

Figure 12-48 shows a block diagram of the cruise control system with PID control added to the system. As is indicated in the block diagram, PID control has a proportional gain, in this case, A; a derivative function with its associated gain, B; and an integral function with an associated gain, C.

As one might expect from the previous discussion on controllers, a PID controller combines the positive effects of all of the above controllers. More specifically, it improves the transient response due to the derivative term,

TABLE 12-3 *Summary of controller performance*

Controller Type	Effect on Transient Response	Effect on Steady State Error
Proportional	Limited improvement	None
PD	Improved	None
PI	Degraded	Improved
PID	Improved	Improved

and reduces or eliminates steady state error due to the integral term.

Table 12-3 is a summary of the four types of controllers discussed.

Composite Example

EXAMPLE 12-10: Given a block diagram of an antenna control system, shown in figure 12-49, with

$$G(s)=\frac{4}{s^2+0.8s+4}$$

find K_{amp}, K_{pos}, and K_{rate}, such that the system meets the following specifications:

$t_p = 1$ sec
$t_{pos} = 4$ sec
$\theta_{ss} = 8$ radians for a 2-volt step input

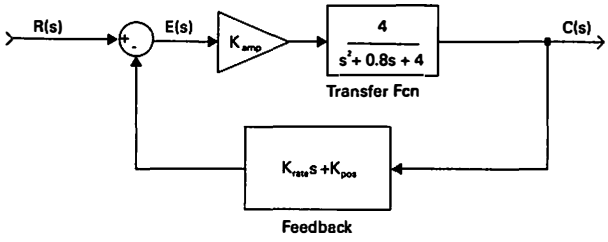


FIGURE 12-49 *System block diagram for Example 12-10*

SOLUTION: The first step is to find the damping ratio and the natural frequency for the overall system given the parameters t_s and t_p:

Using settling time,

$$t_s=\frac{4}{\zeta\omega_n}=4 \text{ sec}$$

$$\zeta\omega_n=1$$

$$\zeta=\frac{1}{\omega_n}$$

Using time to peak,

$$t_p=\frac{\pi}{\omega_n\sqrt{1-\zeta^2}}=1 \text{ sec}$$

$$\omega_n\sqrt{1-\zeta^2}=3.14$$

Squaring both sides gives:

$$\omega_n^2(1-\zeta^2)=9.9$$

Substituting gives:

$$\omega_n^2\left(1-\frac{1}{\omega_n^2}\right)=9.9$$

$$\omega_n^2-1=9.9$$

$$\omega_n=3.3$$

and,

$$\zeta=0.303$$

The next step is to find the CLTF for the overall system:

$$\text{CLTF}=\frac{\frac{4K_{amp}}{s^2+0.8s+4}}{1+\frac{4K_{amp}(K_{rate}s+K_{pos})}{s^2+0.8s+4}}$$

$$=\frac{4K_{amp}}{s^2+(0.8+4K_{amp}K_{rate})s+(4+4K_{amp}K_{pos})}$$

Recalling the standard form for a closed-loop transfer function,

$$\text{CLTF}=\frac{\text{numerator}}{s^2+2\zeta\omega_n s+\omega_n^2}$$

means the following two equations can be written:

$$2\zeta\omega_n=0.8+4K_{amp}K_{rate}$$

$$\Rightarrow 2(0.303)(3.3)=2=0.8+4K_{amp}K_{rate}$$

$$\Rightarrow K_{amp}K_{rate}=0.3$$

and,

$$\omega_n^2=4+4K_{amp}K_{pos}$$

$$\Rightarrow 10.89=4+4K_{amp}K_{pos}$$

$$\Rightarrow K_{amp}K_{pos}=1.72$$

$$\Rightarrow K_{pos}=\frac{1.72}{K_{amp}}$$

At this point, there are two equations, but three unknowns: K_{amp}, K_{rate}, and K_{pos}. The third equation is obtained from the steady state response using the final value theorem:

$$\theta_{ss} = \lim_{s\to 0} s\frac{2}{s}\frac{4K_{amp}}{s^2+(0.8+4K_{amp}K_{rate})s+(4+4K_{amp}K_{pos})}$$

$$\Rightarrow \theta_{ss} = \frac{8K_{amp}}{4+4K_{amp}K_{pos}} = 8$$

$$\Rightarrow K_{amp} = 4+4K_{amp}K_{pos}$$

Substituting from the previous equation set gives:

$$K_{amp} = 4+4K_{amp}\frac{1.72}{K_{amp}} = 10.82$$

Now solving for K_{pos} gives:

$$K_{pos} = \frac{1.72}{K_{amp}} = \frac{1.72}{10.88} = 0.158$$

Finally solving for K_{rate} gives:

$$K_{amp}K_{rate} = 0.3$$

$$K_{rate} = \frac{0.3}{10.88} = 0.0276$$

Summary

A control system is an arrangement of physical components connected or related in such a manner as to command, direct, or regulate itself or another system. Systems themselves can be mathematically modeled by representing the system components with a differential equation. These differential equations can be solved for the unknown variable in the time domain, or converted to the Laplace domain and then solved. The Laplace domain is more frequently used to model systems since mathematical operations are less complicated and since each individual component of the system is more easily expressed.

A transfer function is a mathematical representation of the system and is defined as the ratio of the system output divided by the system input. Knowing the system transfer function and the input, the output is calculated in the Laplace domain by multiplying the input by the transfer function. The output is also referred to as the system response. If this response is converted back to the time domain, a plot known as the time response can be made showing the system response as time varies from zero to infinity.

In viewing the time response plots, six basic types of responses are possible: overdamped, critically damped, underdamped, undamped, unstable oscillating, and unstable non-oscillating. Only overdamped, critically damped, and underdamped systems converge to a steady output value.

Feedback is defined as the comparison of the system response or output with the system input. Open-loop systems are systems without feedback, and closed-loop systems are systems with feedback. There are different types of feedback. If the output signal is fed back without modification, this is referred to as unity feedback. If the signal is altered by a gain, it is called position feedback. Furthermore, the derivative of the output signal can be taken and multiplied by a gain and then fed back to the summing junction. This is called rate feedback.

The difference between output and input is known as the error signal. Ideally, this error signal is zero under steady state conditions, but in cases where it is not zero, the system is said to have steady state error.

The four types of controllers or compensators discussed in this chapter are: proportional control, proportional-derivative (PD) control, proportional-integral (PI) control, and proportional-integral-derivative (PID) control. A summary of their performance is in table 12-3.

Control systems with feedback are essential to the operation of modern weapon systems. They enable speed, precision, and capabilities that can give us advantages from reduced personnel requirements to accurate placement of weapons on target.

---

[i] Smith, 1–3

[ii] Nise, 5

[iii] Sloan

[iv] F414-GE-400, 233

[v] FAS

13 Ballistics and Fire Control

Introduction

The fire control problem may be stated as follows:

> How must a weapon be fired at a stationary or moving target from a launcher that is stationary or moving in order to hit the target?

The first step toward finding the solution is to gather all relative information on the problem. One of the most critical pieces of information is obtained from detecting a target from the background and then obtaining a track history of that target. Previous chapters have described target detection using radar, electro-optical and sonar systems, which results in the determination of the target's actual position. Target information, however, is only part of the input data necessary for solving the fire control problem.

Because weapon platforms are often moving, stabilization and conversion between unstabilized non-inertial and stabilized inertial reference frames is critical for weapon control. An accurate description of the flight path of the weapon is also necessary if the weapon is to be delivered accurately because weapons, in general, do not follow straight paths to the point of target intercept. Physical phenomena acting on the weapon during its time of flight may also cause it to miss its target if the weapon is aimed directly at the target.

The fire control problem can be divided into two major categories that must be solved:

1. The effect of relative target motion during the time of weapon flight
2. The physical phenomena, collectively called ballistics, that produce a curved weapon trajectory

Solving the fire control problem involves the convergence of the target motion and the weapon trajectory, as illustrated in figure 13-1.

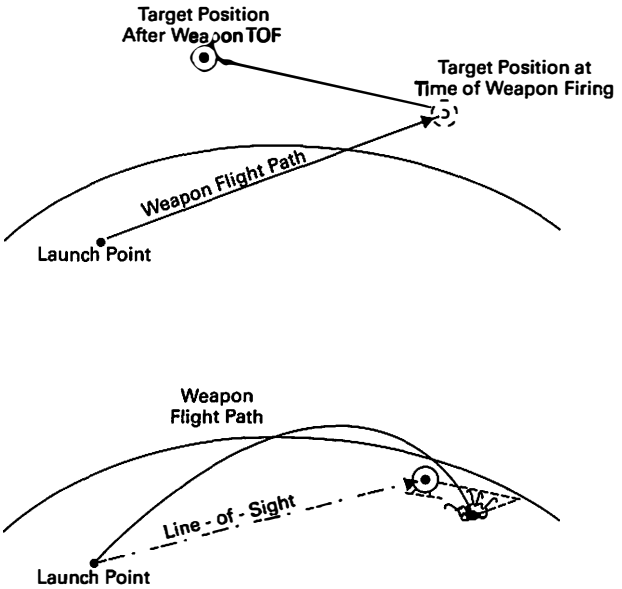


FIGURE 13-1 *(A) Miss-producing effect of a target's relative motion with a straight weapon flight path (B) Miss-producing effect of a curved flight path heading toward a stationary target*

Relative Motion

Relative motion requires aiming a weapon at some point in space displaced from the present target position. The amount of displacement or lead depends on the relative velocity and on the time of flight (TOF) of the weapon from launch point to target. The point of collision between the weapon and the target is based on the future target position. A basic assumption made when computing relative motion is that the target will not accelerate in any direction during the weapon's time of flight. By following a straight-line path, the target simplifies the problem of prediction and thus lowers the target's chances of survival. It might be presumed that a target would avoid this path and use diversionary tactics such as zigzags,

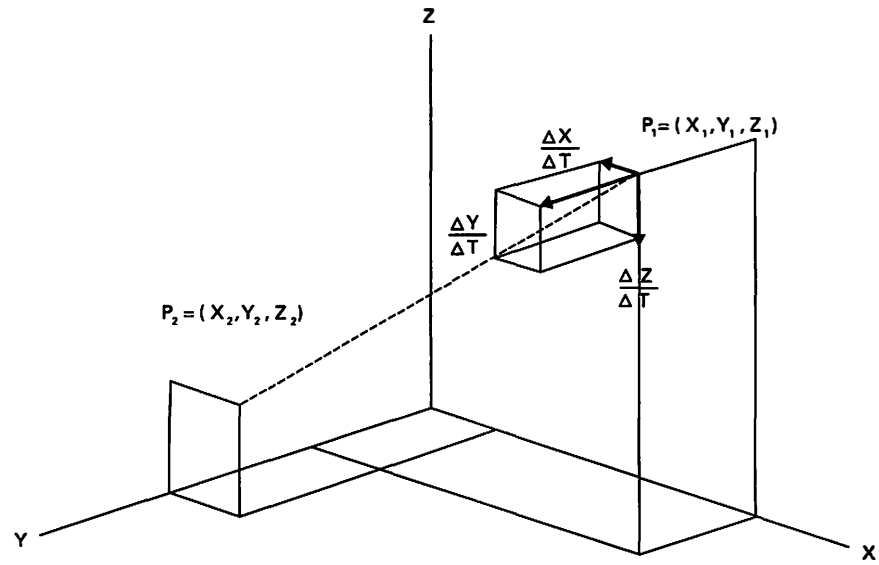


FIGURE 13-2 *Relative motion in Cartesian coordinates*

curved paths, and so on. Certain factors oppose these choices of target paths, however.

The first is that the shortest distance between two points is a straight line; therefore, a target following a straight-line course will be exposed to danger for a shorter period of time. Also, most targets tend to follow a straight-line path. It must be noted that an attacking vehicle must press home its attack. If our presence and the fact that we are prepared to launch a counterweapon causes an attacker to make such radical maneuvers that our weapon would miss, then there is a high probability that their weapon will miss also. Therefore, even if evasive maneuvers are taken by a target, these maneuvers will average over time to a straight-line course or one with a slight curve.

Secondly, it should be realized that even if the target accelerates and/or turns, the shorter the TOF, the less the error. With the above assumptions in mind, the relative motion calculation is reduced to a simple linear function of time:

$$P_2 = P_1 + VT \tag{13-1}$$

where:

P_2 = future position of target
P_1 = present position of target as measured by the sensor subsystem
V = total relative target velocity
T = time of flight of the weapon

As figure 13-2 illustrates, this relationship in a stabilized reference frame actually must be expressed in each dimension:

$$X_2 = X_1 + \frac{\Delta X}{\Delta T}T \tag{13-2}$$

$$Y_2 = Y_1 + \frac{\Delta Y}{\Delta T}T \tag{13-3}$$

$$Z_2 = Z_1 + \frac{\Delta Z}{\Delta T}T \tag{13-4}$$

The future target position derived from the above calculations generally comprises the most significant components of the lead angle. Final aiming angles for the weapon are bearing and elevation, which include target motion effects and corrections as a result of trajectory curvature. The calculation of relative motion solves only part of the weapon control problem by answering the question:

> For a known weapon time of flight, what is the position of the target at the end of that time of flight?

Only when the target's position is known can the effects of weapon-path curvature be computed.

Interior Ballistics

The study of the forces and motion of a projectile as it accelerates inside a gun barrel is called interior ballistics. This applied science includes the combustion of powder,



FIGURE 13-3 *Five-inch shell exiting a naval gun, the end result of interior ballistics*

pressure developed within the cylinder, variations in pressures and velocities, and erosion of the bore. When a charge is ignited inside a closed cylinder, gases form on the surface of each grain of propellant and pressure in the chamber rises rapidly. Due to friction, the projectile does not move until the pressure behind it reaches several hundred pounds per square inch.

For a time after the projectile starts to move, gases are generated more rapidly than the rate at which the volume behind the projectile is increased, and pressure continues to rise. As the propellant-burning area decreases, pressure falls; however, the projectile will continue to accelerate as long as there is a net force acting on it.

When the projectile reaches the muzzle, gas pressure continues to act on the projectile for a short distance beyond the muzzle, and the projectile continues to accelerate for a short time. Figure 13-4 shows the relationship of pressure and velocity versus distance of projectile travel.

Since work performed on the projectile is a function of the area under the pressure travel curve, the muzzle velocity produced by the propellant can be determined using the following familiar equation.

$$\text{work} = \text{KE} = \frac{1}{2}mv^2, \text{ if inital velocity is zero} \quad (13\text{-}5)$$

where:

v = muzzle velocity (m/s)
m = mass of projectile (kg)

Should it be desired to increase the muzzle velocity of a projectile, the work performed or the area under some new curve must be greater than the area under a curve at a lower muzzle velocity. Such an increase in velocity is indicated by curve B, whose maximum pressure is equal to that of curve A, but whose area is greater than that under curve A. It appears that the ideal pressure travel curve would be one that coincides with the curve of maximum permissible pressure.

Even if it were possible to design such a propellant, it would be undesirable to do so. In addition to producing excessive erosion (which would decrease the accurate life of the gun), brilliant flashes and non-uniform velocities due to high muzzle pressure would result. The powder chamber would also have to be enlarged, thus increasing the weight and decreasing the mobility of the gun.

Curve C is the gun strength curve, which indicates maximum permissible pressure and is a function of gun tube thickness. There are several methods available for changing the pressure travel curves of a given gun. The easiest method is to change the type of propellant grain.

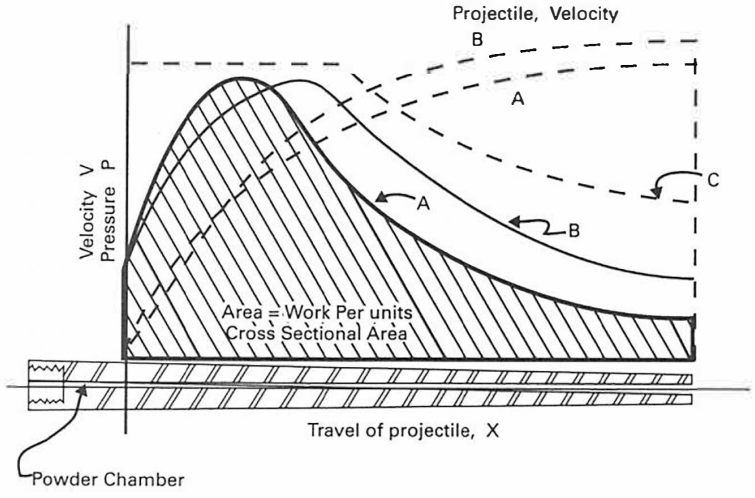


FIGURE 13-4 *Interior ballistics diagram showing pressure travel (solid lines) and velocity travel (dotted lines) curves. Curve C is maximum pressure verse position.*

Note: Often the pressure travel curve is the given variable, and the gun tube thickness is then dictated by the pressure-travel curve.

Propellant Types

All guns and most rocket-powered weapons use solid propellants to provide their propulsion. The first solid propellant used by man was black powder in the thirteenth century. Black powder is no longer considered suitable for use as a propellant for several reasons: it burns incompletely, leaving large amounts of residue; it creates high temperatures, resulting in rapid bore erosion; it creates great billows of black smoke; and it detonates rather than burns.

Gunpowders or smokeless powders are the propellants in use today. These substances are produced by combining nitrocellulose (nitric acid and cotton) with ether and alcohol to produce a low explosive. Although they are called smokeless powders, they are neither smokeless nor a powder, but in granule form.

Solid propellants are designed to produce a large volume of gases at a controlled rate. Gun barrels and some rocket casings are designed to withstand a fixed maximum gas pressure. The pressure generated can be limited to this maximum value by controlling the burning rate of the propellant. The burning rate is controlled by varying the following factors:

1. Size and shape of the grain, including perforations.
2. Web thickness or amount of solid propellant between burning surfaces; the thicker the web, the longer the burning time.
3. Burning rate constant, which depends on the chemical composition of the propellant.
4. Percentages of volatile materials, inert matter, and moisture present. A 1% change in volatiles in a low-volatile-content propellant may cause as much as a 10% change in the rate of burning.

When a propellant burns in a confined space, the rate of burning increases as both temperature and pressure rise. Since propellants burn only on exposed surfaces, the rate of gas evolution or changes in pressure will also depend upon the area of propellant surface ignited. Propellant grains are classified according to variations in their rates of burning due to changes in the amount of ignited surface; three classifications of burn rates are the following:

1. Degressive—as burning proceeds, the total burning surface decreases. Propellants formed into pellets, balls, sheets, strips, or cord burn degressively. See figure 13-5.

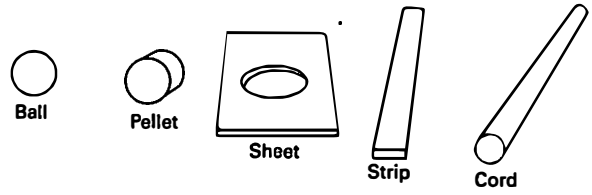


FIGURE 13-5 *Degressive burning grains*

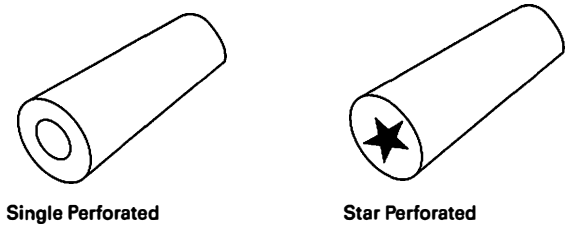


FIGURE 13-6 *Neutral burning grains*

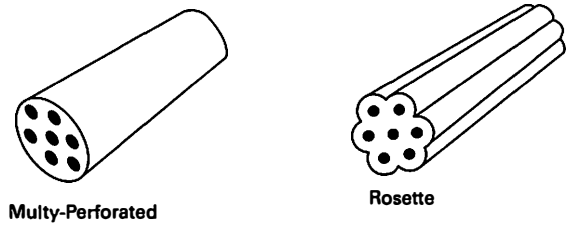


FIGURE 13-7 *Progressive burning grains*

2. Neutral—as burning proceeds, the total burning surface remains approximately constant. Single perforated grains and star perforations result in neutral burning. See figure 13-6.
3. Progressive—as burning proceeds, the total burning surface increases. Multiperforated and rosette grains burn progressively. See figure 13-7.

Figure 13-8 illustrates the effects that the three types of powder grains have on the pressure travel curve. As shown, a progressive powder reaches maximum pressure much later than a degressive powder and loses pressure much more gradually. The degressive powder reaches a much larger maximum pressure sooner. Thus, it stands to reason that degressive powders are used in weapons with short tubes, and increasingly progressive-type powders are used in weapons with longer tubes.

Other methods of changing the pressure travel curve involve altering the burning rate of the powder, changing the loading density of the propellant, or changing the grain size.

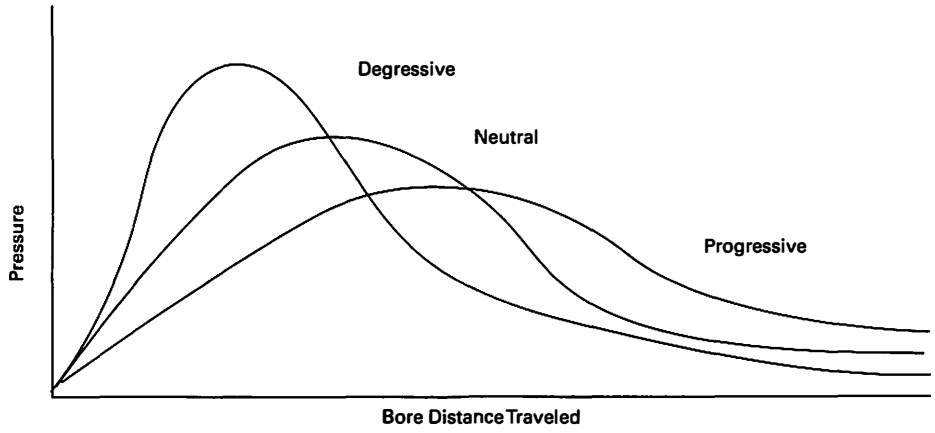


FIGURE 13-8 *Pressure travel curve*

The energy developed by burning propellant in a typical medium-caliber gun, assuming complete combustion, is estimated to be distributed as shown in table 13-1:

TABLE 13-1 *Distribution of propellant energy*

Energy Absorbed	Percent of Total
Translation of projectile	32.00
Rotation of projectile	0.14
Frictional work on projectile (due to engraving of rotating bands, wall friction, and effects of increasing twist)	2.17
Translation of recoiling parts	0.12
Translation of propellant gases	3.14
Heat loss to gun and projectile	20.17
Sensible and latent heat losses in propellant gases	42.46

Rifling

Projectiles are usually spun as they proceed down the barrel to provide aerodynamic stability in flight. The cause of the spin in the barrel is called rifling. Rifling is either indentations or projections that twist inside the length of the barrel. The raised portions of the rifling are called lands, and the lower surfaces are called groves. The bullet's soft rotating band engages the rifling, which causes the spin.

Muzzle Velocity

Muzzle velocity, sometimes called initial velocity, is the velocity of the projectile immediately after leaving the bore. It is desirable to maintain muzzle velocity at a constant value for each round, thereby determining range and accuracy of the gun system. Variations in the structure of the gun and projectile as a function of size and characteristics of powder charge affect the pressure and muzzle velocity. The following variables are direct contributors to muzzle velocity.

1. *Gun tube length.* Increasing the caliber or length of the gun tube increases muzzle velocity. This is clearly indicated in pressure travel curves. Since energy from the expanding gases remains after the projectile leaves the weapon, lengthening the gun tube enables more of this energy to be coupled to the projectile, thus increasing its muzzle velocity. However, lengthening the gun tube increases the weight and thus decreases the weapon's mobility. Therefore, beyond a certain point, the added velocity is not worth the additional weight.
2. *Projectile weight.* Increasing projectile weight requires a greater force to move it. Thus, maximum pressure is greater and is reached earlier in the projectile's travel and causes a lower muzzle velocity.
3. *Temperature of powder.* Range tables are based on a powder temperature of 90°F at the time of firing. Increasing this temperature increases the burning rate and the maximum pressure of the propellant, resulting in greater muzzle velocity. In a 5-inch gun, for example, the muzzle velocity will be increased approximately 2 feet per second for each degree of increase in temperature; a decrease in temperature will cause a corresponding decrease in velocity.

4. *Temperature of gun.* Rapid firing produces a high temperature, which, if allowed to act on a round for any length of time, can have an appreciable effect. If the gun is hot enough and the round is allowed to remain unfired in the breech long enough, spontaneous firing occurs. This is commonly known as cook off.
5. *Gun bore erosion.* The accuracy of a gun depends on the condition of its bore, which is determined by the degree of wear. There are two types of wear: erosion—the result of firing and corrosion—the result of neglect. Erosion is greatest at the origin of rifling, where the temperature of the propelling gases is highest. The process of erosion wears down the lands, causing the bore to become irregular and enlarged. The end result is that erosion causes loss of pressure and, subsequently, muzzle velocity, which ultimately enlarges the bore to the extent that spin is no longer imparted to the projectile. All erosion factors are related to the temperatures of the expanding gases and the length of time these gases are present in the bore. Therefore, large guns with slower powders and larger barrels undergo more erosion per round fired than small guns.

Erosion can be controlled but not eliminated and eventually renders the gun ineffective or unsafe. The signs of erosion are progressive losses in accuracy and muzzle velocity. For each class of gun, there are curves that show the relationship between enlargement of the bore and muzzle velocity of the gun. Thus, by checking bore diameter, velocity loss can be predicted and compensated for. In addition, barrel liners can be replaced before the performance of the gun becomes noticeably erratic.

A special measuring gage in the form of inside calipers that measure the inside diameter at several points along the length of the barrel, and from this data muzzle velocity can be predicted. Since this gage is not used after each firing, curves relating the number of rounds fired to loss in muzzle velocity are used to estimate the amount of erosion.

Exterior Ballistics

Exterior ballistics is a generic term used to describe a number of natural phenomena that tend to cause a weapon in flight to change direction, velocity, or both. These phenomena are gravity, drag, wind, drift, and Coriolis effect.

Computation of relative motion effects produces the actual point of collision between the weapon and the target. The computation of the effects of exterior ballistics produces aiming corrections as a result of the curvature of the weapon's flight path. These corrections don't change the final collision point, but cause the trajectory of the weapon to intersect the flight path of the target.

The ultimate aim of ballistic computation, then, is to produce angular corrections to bearing and elevation values derived from relative motion results. Inherent in these calculations is the computation of the time of flight of the weapon, which in turn is used to facilitate the relative motion computation.

Ideally, the above computations would completely describe the complex trajectory of a weapon. When using coordinates of the future target position as input data, these computations would produce angles of launch and weapon time of flight as output, as depicted in figure 13-9. It must be emphasized that the central problem of weapon delivery is finding the solution of the intersection of two curves, one described by the motion of the target and the other described by the motion of the weapon.

Gravity

The force of gravity always tends to accelerate the movement of an object downward toward the center of the earth.

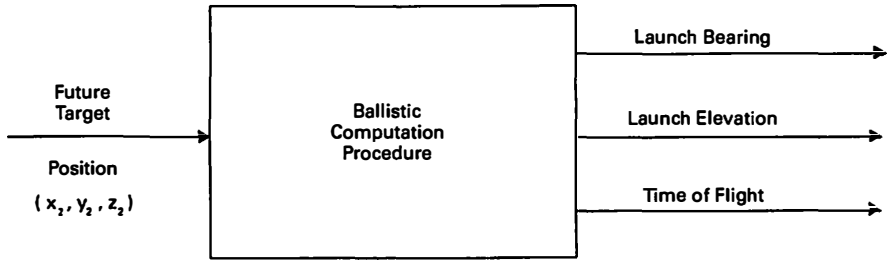


FIGURE 13-9 *Input/output relationship of a ballistic procedure*

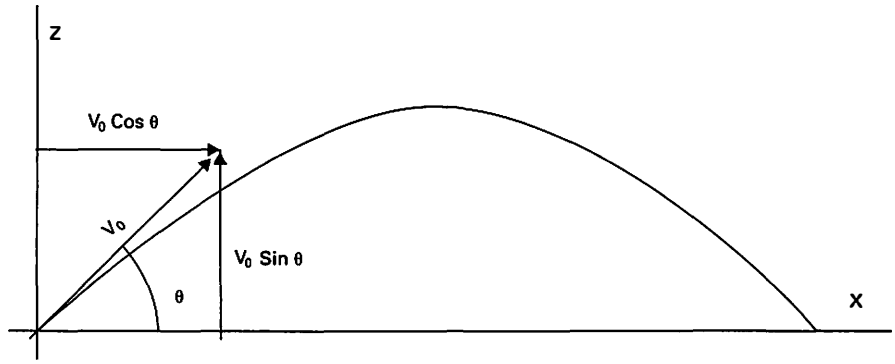


FIGURE 13-10 *Weapon trajectory that considers gravity alone acting on the weapon*

Effects of this force on the trajectory of a weapon are well known and, if considered alone, result in a very simple solution to the trajectory problem.

Figure 13-10 is an illustration of the trajectory of a weapon in a vacuum and the effect of gravity upon that weapon. In a vacuum on a "flat" nonrotating earth, the horizontal component of initial velocity ($V_0 \cos\theta$) remains unchanged. However, the vertical component ($V_0 \sin\theta$) is continuously changing as a result of the acceleration due to gravity. This changing vertical velocity results in a parabolic trajectory.

A ballistic procedure that provides an approximation to the desired output is developed from equations that describe this trajectory. A two-dimensional example will serve to illustrate this procedure.

The differential equations of motion are:

$$\frac{d^2x}{dt^2}=0 \tag{13-6}$$

$$\frac{d^2z}{dt^2}=\frac{dz'}{dt}=z''=-g \tag{13-7}$$

subject to the initial conditions:

$$x(t=0)=0 \tag{13-8}$$

$$x'(t=0)=v_0\cos\theta \tag{13-9}$$

$$z'(t=0)=v_0\sin\theta \tag{13-10}$$

$$z(t=0)=0 \tag{13-11}$$

These are solved by separating variables, and integrating and evaluating constants of integration (through the use of initial conditions). Thus, from equation 13-6:

$$\frac{dx}{dt}=x'=v_0\cos\theta \tag{13-12}$$

$$\int dx=v_0\cos\theta\int dt=x=(v_0\cos\theta)t+C_1 \tag{13-13}$$

Evaluate the constant of integration, C_1:

$$x(0)=0+C_1;\ \text{hence,}\ C_1=0\ \text{and}$$
$$x=tv_0\cos\theta\ \text{ or }\ t=\frac{x}{v_0\cos\theta} \tag{13-14}$$

Solve equation 13-7 for z:

$$dz''=-gdt;\ \text{hence,}\ z'=-gt+C_2 \tag{13-15}$$

Evaluate the constant of integration, C_2:

$$z'(0)=0+C_2=v_0\sin\theta;\ \text{hence,}\ C_2=v_0\sin\theta$$
$$z'=\frac{dz}{dt}=-gt+v_0\sin\theta \tag{13-16}$$

Solve for z:

$$\int dz=\int(-gt+v_0\sin\theta)dt$$
$$z=\frac{gt^2}{2}+t(v_0\sin\theta)+C_3 \tag{13-17}$$

Evaluate the constant of integration, C_3:

$$z(0)=0+0+C_3=0;\ \text{hence,}\ C_3=0$$
$$z=-\frac{g}{2}t+(v_0\sin\theta)$$

or

$$\theta=\arcsin\left(\frac{z+0.5gT^2}{v_0T}\right) \tag{13-18}$$

The procedure for the solution of equations 13-14 and 13-18 is one of successive approximations. The concept

involves computing a first approximation to t by using equation 13-14 and $\theta = \arctan(z/x)$. Next compute the effect of gravity on θ using equation 13-18. Use the results of equation 13-18 to compute a new value for t. Continue this alternate substitution of variables until the values of t and θ converge to a final value that is accurate enough for use.

This procedure is common to all fire control systems and must be well understood. Sufficient accuracy can be determined by testing the absolute difference between t_n and t_{n-1} as it approaches zero, where n is the number of iterations. Figure 13-11 is an annotated flow diagram of this procedure.

The above ballistic procedure is valid regardless of the elevation of the target's future position, since equations 13-14 and 13-18 describe the trajectory that passes through the point (x, z), which defines this position. This validity (in a vacuum) can be shown by comparing the trajectories of a surface target ($z = 0$) and an air target ($z > 0$). Refer to figure 13-12.

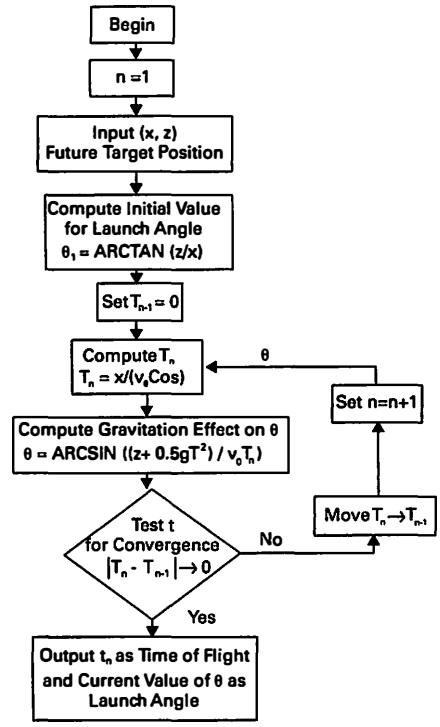


FIGURE 13-11 *Flow diagram for the ballistic procedure considering gravity only*

Drag

In the vacuum trajectory situation, the solution was easily obtained because of simplifications and assumptions made in formulating the problem. When the effects of the atmosphere are considered, forces other than gravity act on the weapon, causing it to deviate from a pure parabolic path. Forces that act opposite to the projectile's motion are called drag. There are three types of drag force, which apply to projectiles.

1. Skin drag is caused by friction on the projectile's outer surface as it moves through the air, resulting in heat.
2. Shape drag is a low-pressure zone produced behind the projectile due to the flow of air around the projectile. This low pressure tends to retard forward motion.

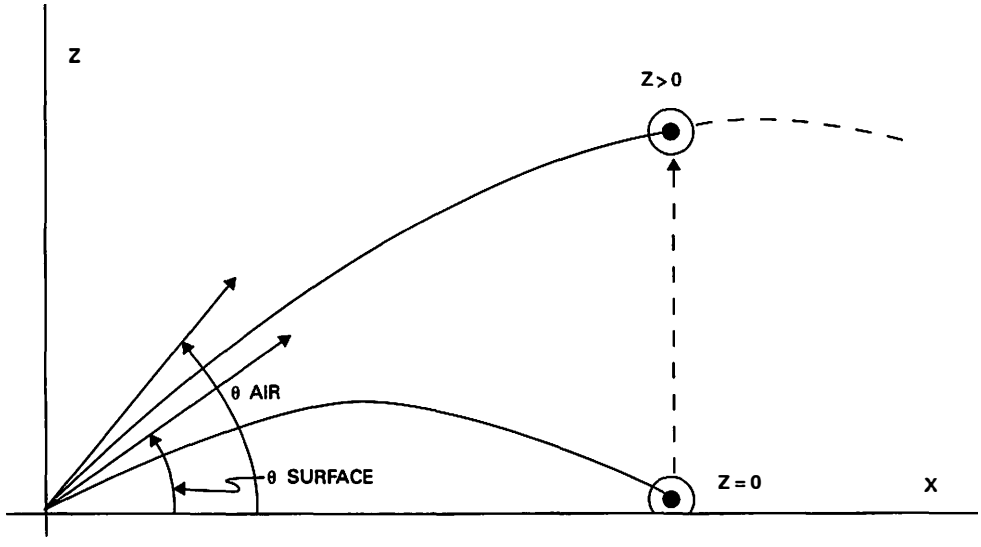


FIGURE 13-12 *Comparison of surface target and air target trajectories*

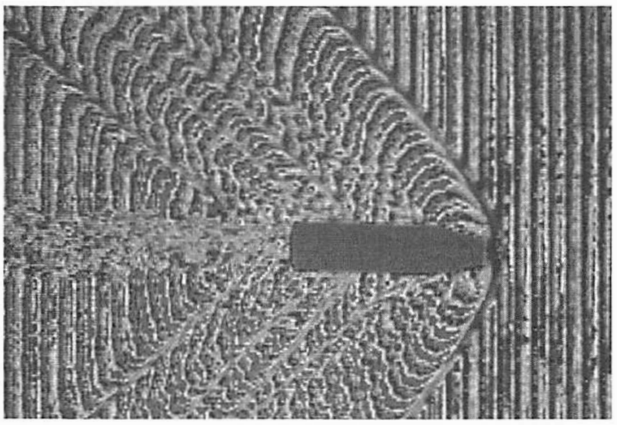


FIGURE 13-13 *Airflow around a 20-millimeter shell flying at about 500 m/sec (about 1,600 ft/sec), showing the three types of drag. The image was recorded using polarized light with an exposure time of 20 nanoseconds.*

3. Wave drag transforms projectile energy into acoustic waves as the body moves through the air. This effect is of particular importance near the speed of sound in air. See figure 13-13.

The combined effect of energy loss is a deceleration of the projectile during its time of flight. This deceleration acts to decrease the velocity of the weapon as it travels through the atmosphere. For a given projectile form, retardation in velocity is directly proportional to atmospheric density and to velocity itself.

This deceleration, called parasitic drag (D), is given by the following relationship:

$$D = C_D \times A \times \frac{1}{2}\rho v^2 \qquad (13\text{-}19)$$

where:

C_D = drag coefficient derived from empirical data
A = cross-sectional area of the projectile
ρ = atmospheric density that is a function of altitude and atmospheric temperature at a given height
v = velocity of the projectile at any instant of time during its flight

Parasitic drag always acts in a direction opposite to that of the weapon's motion, as shown in figure 13-14. As a result, two fundamental interdependent actions take place while a weapon is in flight. First, as illustrated in equation 13-19, the force of drag at any point in the weapon's trajectory is a direct function of the velocity at that point. Because drag forces are continuously retarding the velocity of the weapon, the force of drag itself is continuously changing over the entire time of flight of the weapon, with drag dependent upon velocity and velocity in turn dependent upon drag. Thus, the magnitude of both velocity and drag is changing, with each being dependent upon the other.

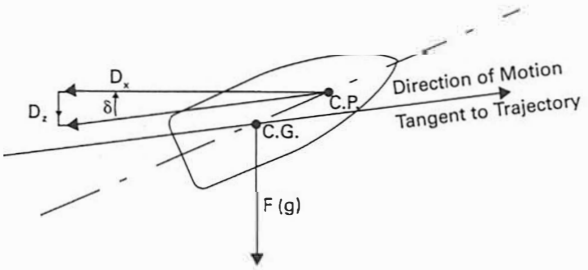


FIGURE 13-14 *Parasitic drag and gravity acting on a projectile in flight. The relative positions of the center of gravity (C.G.) and the center of pressure (C.P.) are dependent upon the projectile's configuration.*

Second, as the weapon travels the curved path of the trajectory, the angle δ is never constant, which results in the force of drag continuously changing direction throughout the entire flight of the weapon. Therefore, the force of drag constantly changes in both magnitude and direction.

The end result of drag is a slowing of the projectile over its entire trajectory, producing a range that is significantly shorter than that achievable in a vacuum. This effect is shown in figure 13-15, over a range of drag coefficients.

A discussion concerning the difference between the vacuum and atmospheric trajectories is warranted. Whereas the vacuum trajectory is symmetric about the

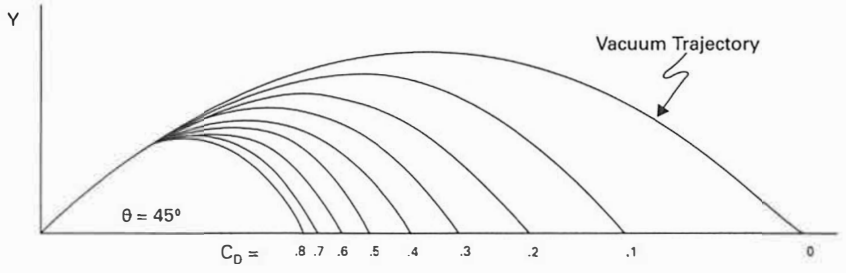


FIGURE 13-15 *Comparison of vacuum and atmospheric trajectories*

mid-range value, with angle of fall equaling angle of departure, the atmospheric trajectory is asymmetric. The summit of the trajectory is farther than mid-range, and the angle of fall is greater than the angle of departure.

Additionally, the atmospheric trajectory is dependent on projectile mass due to external forces, which, unlike for the vacuum trajectory, cause the atmospheric trajectory's horizontal component of velocity to decrease along the flight path. This results in a computation of the necessary departure angle needed to affect maximum range, whereas in a vacuum the maximum range's departure angle is 45°.

When considering the elevation at which maximum range is achieved, there are two offsetting effects: the change in drag with projectile velocity and atmospheric density. A projectile's velocity is greatest immediately upon leaving the gun; thus drag is also greatest at this time.

For small, lightweight projectiles, deceleration is greatest at the beginning of flight; therefore, to take advantage of the improvement in horizontal range while the velocity is greatest, it is necessary to shoot at elevations less than 45°.

In large projectiles it is best to employ elevations above 45° to enable the projectiles to quickly reach high altitudes where air density and thus drag are decreased. Range increases for large projectiles result from the approximation of vacuum trajectory conditions in the high-altitude phase of projectile flight.

Wind

The effects of wind can be divided into two components: range wind and cross wind. Range wind is that component of wind that acts in the plane of the line of fire. It serves to either retard or aid projectile velocity, thus either increasing or decreasing range. Cross wind acts in a plane perpendicular to the line of fire and serves to deflect the projectile either to the right or to the left of the line of fire. Wind acts on the projectile throughout the time of flight; therefore, the total deviation in range and azimuth is a function of the time of flight.

Drift

Drift of a gun projectile is defined as "lateral displacement of the projectile from the original plane of fire due to only the effect of rotation of the projectile." The principal cause of drift lies in the gyroscopic properties of the spinning projectile. According to laws of gyroscopic motion, the projectile seeks, first of all, to maintain its axis in the direction of its line of departure from the gun. The projectile's center of gravity, however, follows the curved path of the trajectory, and the instantaneous direction of its motion at any point is tangent to the trajectory at that point. Therefore, the projectile's tendency to maintain the original direction of its axis results in the projectile's leaving this axis pointed slightly above the tangent to the trajectory.

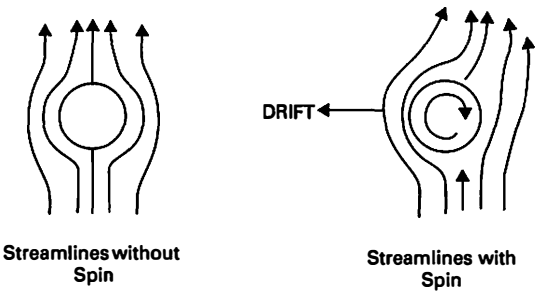


FIGURE 13-16 *Magnus effect*

The force of the air resistance opposed to the flight of the projectile is then applied against its underside; see figure 13-16. This force tends to push the nose of the projectile up, upending it. However, because of gyroscopic stabilization, this force results in the nose of the projectile going to the right when viewed from above. The movement of the nose of the projectile to the right then produces air resistance forces that tend to rotate the projectile clockwise when viewed from above. However, again because of gyroscopic stabilization, this force results in the nose of the projectile rotating down, which then tends to decrease the first upending force. Thus, the projectile curves or drifts to the right as it goes toward the target. This effect can easily be shown with a gyroscope.

A more detailed analysis of gyroscopic stability would show that a projectile is statically unstable, but because of the gyroscopic effects, it becomes dynamically stable. These effects actually cause the axis of the projectile to make nutations about its flight path. Also, the Magnus effect, where a rotating body creates lift (figure 13-16), must be considered. The Magnus effect is what causes a pitched baseball to curve.

Coriolis Effect

The preceding effects have all considered the earth to be flat and nonrotating. Of course, this is not true of the earth in real life, and Coriolis acceleration caused by the rotating earth must be considered for projectiles traveling great distances. Centripetal acceleration of a point on the earth's surface will vary with its distance from the rotation axis. The difference in the rotational velocities of the

launcher and the target due to their different locations on the earth's surface also must be considered.

Coriolis acceleration is created when a body or projectile moves along a radius from the rotation axis of the earth (or a component of it). The amount of rotation (also known as the earth rate), which is dependent on the latitude, is given as:

$$\text{earth rate} = (2\pi \text{ radians})/(24 \text{ hours}) \times \sin(\text{lat}) \quad (13\text{-}20)$$

As the frame of reference moves under the projectile that is traveling in a straight line, it appears to be deflected in a direction opposite to the rotation of the frame of reference. Thus, this acceleration tends to curve an object to the right in the Northern Hemisphere and to the left in the Southern Hemisphere.

An observable weather-related example would be that the air moving outward from a high-pressure region becomes clockwise wind, and air moving into a low-pressure region becomes counterclockwise wind.

EXAMPLE 13-1: Consider a projectile traveling north at 700 m/s at a latitude of 40° N. What is the deflection from the aim point caused at 25 kilometers, neglecting all other forces and assuming a 35-second time of flight?

Solution: The acceleration caused by Coriolis is:

At 40° N, the earth rate is 0.17 rad/hr (4.7×10^{-5} rad/sec).

$$\text{acceleration}_{\text{Coriolis}} = 2 \times \text{velocity} \times \frac{2\pi}{24 \text{ hr}(3{,}600 \text{ s/hr})} \times \sin(\text{lat})$$

$$= 2(700 \text{ m/s})\left(\frac{2\pi}{86{,}400 \text{ s}}\right)(\sin 40) = 0.065 \text{ m/s}^2$$

$$\text{displaced distance} = \frac{1}{2}\text{acceleration}_{\text{Coriolis}} \times \text{time of flight}^2$$

$$= \frac{1}{2}(0.065 \text{ m/s}^2)(35 \text{ s})^2 = 39.8 \text{ meters}$$

Therefore, for long-range unguided projectiles, the Coriolis correction is very important.

Ballistics Procedure

The ballistics procedure described earlier is a highly simplified illustration of the concept of trajectory calculations. The calculations involved in a practical trajectory procedure are extremely complex and must take into account factors such as drag, lift, drift, and wind as well as the force of gravity.

The two basic equations 13-14 and 13-18 must be expanded to include all of the forces acting on the weapon over its time of flight. Although a practical ballistic procedure is very complex in detail, the logic flow is identical to that described in figure 13-11. The output answers the second major question in the solution of the fire control problem:

> For a known point in space, what will be the launch angle and the time of flight of a weapon to that point?

Prediction Calculation of Future Target Position

Now that relative motion and ballistic procedures have been investigated separately, it must be pointed out that in reality, each is dependent upon the other. In the statements of the relative motion and ballistics problems given earlier, it will be noted that the solution for future target position depends on time of flight, and in turn, the solution for the ballistic trajectory and time of flight depends upon future target position.

To illustrate the problem a bit further: For any given point in space, a trajectory and time of flight can be computed, but the correct point in space for intercept cannot be given directly since that point depends upon time of flight for its calculation. In this context, *time of flight* is a generic term for the time of travel of the weapon.

The prediction algorithm described below is general in nature and is directly applicable to all weapon systems. The prediction procedure incorporates both the relative motion procedure and a ballistic procedure in a closed-loop manner. Since at the outset neither the future target position nor the final time of flight is known, the solution is an application of the principle of successive approximations. Since the measured present target position is the first-known point in space, this position can be employed to produce a first approximation to the launch angle and time of flight (see trajectory #1 in figure 13-17).

If the weapon were fired based upon this first approximation, it would obviously pass well behind a moving target. Now compute just how far behind the target the projectile would pass. In other words, where is the target at the end of the time of flight compared to the present position?

Next, compute the new position and a new point. If the weapon were fired at this new point, it would pass well ahead of the target. Using the time of trajectory #2, compute a new target position and a third trajectory and time; see trajectory #3.

Continue computing alternate target positions and times until the difference between the two most recently computed times is very small, say, 10^{-5} second. When this occurs, the point in space used to compute the time of flight will be the same point that results when time of

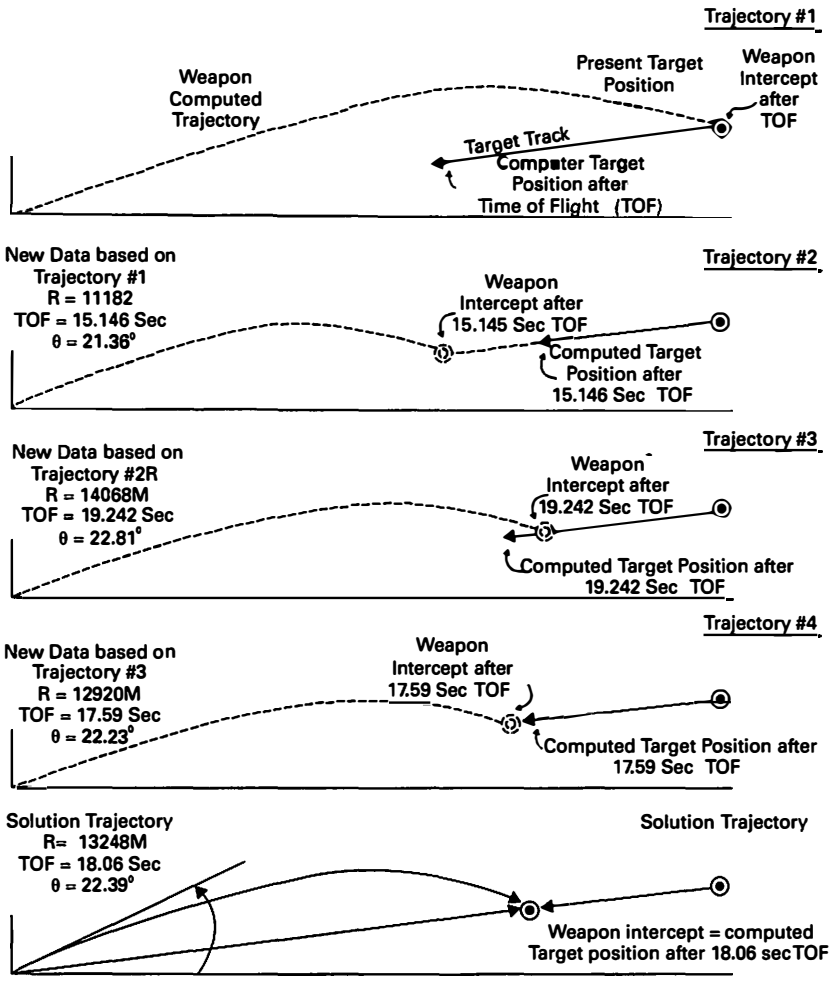


FIGURE 13-17 *Illustration of the iterative procedure employed in a fire control solution; only four of ten iterations are shown, along with the solution trajectory that passes through the future target position*

flight is applied to the relative motion solution; thus, the fire control problem is solved.

Reference Frames and Coordinate Systems

The solution to the weapon control problem involves gathering and processing information relating to position and motion of a target. In order to be useful, position and motion data must be related to a known reference. For example, an aircraft's velocity does not provide any meaningful information unless it is known whether the velocity is relative to the air in which it is flying, to the ground over which it travels, or perhaps even to the distant fixed stars. Frames of reference are characterized by several major properties: an origin, reference lines, reference planes, and stability.

To be useful for measuring linear distances, a reference frame must have a point of origin that establishes the zero point for the reference frame. A distance of 100 yards has meaning within the frame since it is understood that the measured distance is relative to the frame's origin.

Simply measuring distance is not sufficient to establish position within a particular reference frame. In weapon systems, target direction from the origin is measured in angular quantities. To give meaning and a zero reference to angular quantities, reference lines are established. Reference lines pass through the origin and establish the axes of the reference frame.

In addition to reference lines, angular measurements require the definition of planes in which angles are to be measured. These are called reference planes. For example, an object located a distance of 100 yards from the origin

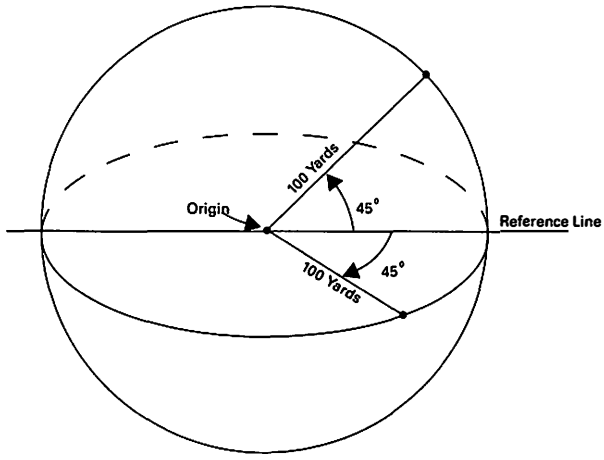


FIGURE 13-18 *The position of an object in two different planes*

and 45° from a reference line has virtually an infinite possibility of actual locations unless the plane in which the angle is being measured is defined. Figure 13-18 illustrates two possible locations based upon a distance and an angle measured in two different planes.

A general reference frame can now be constructed based upon the three properties described above. Since it is desired to express position with both linear and angular quantities, the general reference frame is comprised of three reference axes (X, Y, Z), with the bearing reference line arbitrarily chosen as the X axis. A horizontal reference plane is established as the plane that contains the X and Y axes. A vertical reference plane is also employed and is defined as the plane that contains the target and the Z axis. Figure 13-19 illustrates this general reference frame.

The property of reference frame stability directly affects the description of target motion within the reference frame. A truly stable or inertial reference frame is one

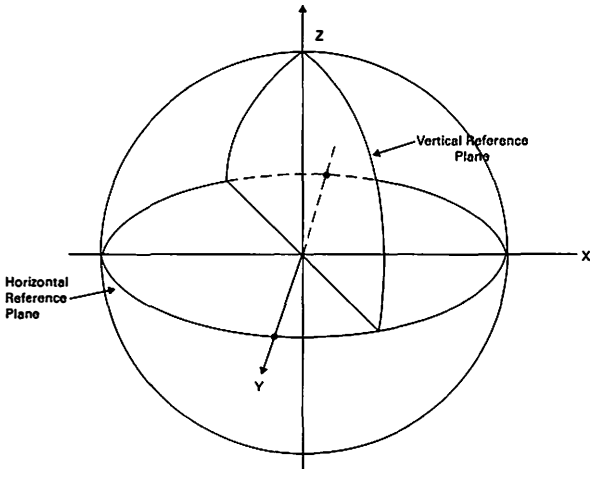


FIGURE 13-19 *General three-dimensional reference frame*

that is held fixed or nonrotating with respect to distant stars. In this type of reference frame, target motion is very easily described since all changes in target position within the frame are due only to motion of the target itself. An unstable or non-inertial reference frame is one that is constantly rotating in a random manner; therefore, changes in target position within such a reference frame are the result of both target motion and reference frame rotation.

A reference frame may have virtually any point in the universe as its origin. However, reference frames useful in weapon control originate either at the center of the earth or at the weapon station itself. In this text, attention will be concentrated on reference frames that originate at the weapon station.

Weapon Station Reference Frames

Weapon station reference frames are fixed to, and move with, weapon platforms such as ships, aircraft, submarines, and controlled missiles. Frames of reference used in weapon control are of two general types: unstable and thus non-inertial, and stabilized and considered inertial for the computation of target motion.

The unstabilized weapon station frame has its origin at one of the sensors of the fire control system. The reference axes are defined by the architecture of the weapon platform. One axis is oriented fore and aft, another is oriented horizontally and perpendicular to the fore-and-aft axis, and a third is oriented vertically, perpendicular to the other two axes.

The major characteristic of an unstabilized reference frame is that it rotates about all three of its axes as the weapon station rotates in pitch, roll, and yaw. For this reason, it is considered non-inertial and thus not suitable for the description of target motion.

The origin of the stabilized weapon station frame is again at the weapon station. However, the reference axes are no longer coincident with the structure of the vehicle. This is due to the incorporation of inertial elements (gyroscopes) within the weapon station, which establish the axes for this frame of reference.

One such element is made to track the axis of the earth's rotation, thus establishing a permanent north-south horizontal reference line. A second horizontal reference line is established by the perpendicular east-west line derived from the same inertial element as the north-south line. Another inertial element is made to track the center of the earth, and thus establish a fixed vertical axis.

Because the axes of this reference frame are fixed to true north and the center of the earth, the stabilized weapon station frame is considered inertial during the flight times of most weapons. For this reason, the stabilized

reference frame is employed for the description of target motion and the computation of lead angles.

Figure 13-20 illustrates that in any given weapon system, there may be several functional reference frames. A surface combatant is used here as an example because it is on this type of weapon station that this concept of multiple reference frames is most apparent.

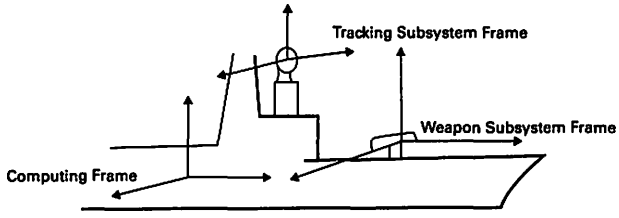


FIGURE 13-20 *Several weapon station frames in a single weapon control system*

The sensor and tracking systems may gather target information in one weapon station frame (unstable) while the computational system computes lead angles in another (stable), and the launching system is positioned in still a third frame (unstable). Note that these related frames of reference are displaced from one another and may rotate with respect to one another.

Gyros

To accurately translate between non-inertial reference frames and inertial reference frames, the three-dimensional motion must be measured. The interaction between stable inertial masses and moving platforms makes gyroscopes ideal for measuring motion or, more specifically, measuring angular differences. The word *gyro* colloquially used today refers to any instrument that accomplishes this measurement.

Gyroscope

Fundamentally, a gyro is a rapidly spinning rotor. Its rotational momentum gives it two useful properties: rigidity in space and precession. Rigidity in space refers to the tendency of a rotor spun on its axis to remain in a fixed attitude if no external force is applied. Precession is the tendency of the spinning rotor axis to turn at right angles to the direction of an externally applied force.[i]

Gyroscopic inertia enables a gyro to remain in a fixed orientation in space, allowing the gyro to serve as a stable directional reference. When positioned orthogonally, a set of three gyros can provide directional reference in the *X*, *Y*, and *Z* planes and measure the roll, pitch, and yaw motions of the vehicle or platform carrying it.

Gyros' rigidity has been used in ships to resist rolling moments from wave action; however, this system requires massive rotors weighing up to 120 tons, which is not a practical means of stabilization.

Modern methods of stabilizing ships, aircraft, and weapon platforms employ comparatively small gyros as directional references, which make them useful for supporting autopilot-type systems. In autopilot systems, relative movement between the platform and the gyro creates an angular displacement, which can be calculated as an error and used to generate control signals.

When connected to hydraulic or electrical control systems, gyros can prompt control surface deflections to correct the orientation of a ship or command a waypoint turn to keep a cruise missile on-course. In each case, the gyro (called a free gyro in this case) is permitted to remain stationary in space and thereby moves a variable resistor as the platform moves about it. Gyros therefore are at the heart of measuring the forces that can move vehicles off-course or (indirectly) measuring distances to keep vehicles on-course.

A gyro's spin axis has a tendency to turn at right angles to the direction of an applied force; see figure 13-21. It is this precession that causes the flight path of spin-stabilized projectiles to curve in the horizontal plane as discussed earlier. As depicted in figure 13-22, when a

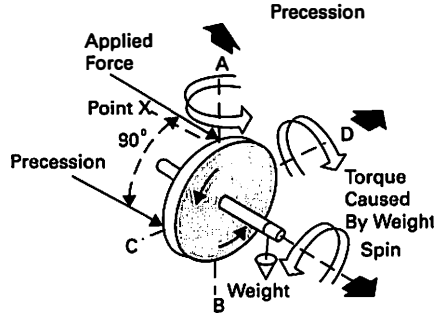


FIGURE 13-21 *Gyroscopic forces*

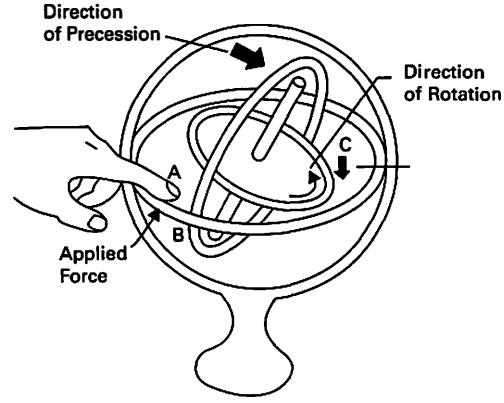


FIGURE 13-22 *Gyro precession*

downward force is applied to the end of the gyro spindle, a counterclockwise precession results. Thus, a force applied in the vertical plane causes the gyro to precess in the horizontal plane. Unlike the free gyro, a rate gyro is fixed in one axis, and the torque of precession is converted into an electrical signal that is used to compute angular displacement.

An integrating gyro combines the inertial rigidity of a free gyro with the torque-measuring features of a rate gyro. Gyro inputs include roll (or pitch angle) and target position angle error, which are combined into a single output. The major input component is the position angle error.

Associated position-error voltages from the sensor vary in a jagged manner. If the drive system were allowed to follow the error voltage signal, the result would be a series of abrupt jumps and a rough positioning of the tracking element. Because of the gyro's inertia, the gyro cannot respond instantly to the changing input signal. This causes a smoothing process that results in an averaging of the rapidly changing input. Figure 13-23 illustrates this electrical input/output relationship.

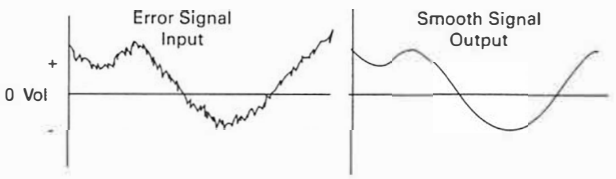


FIGURE 13-23 *Input/output relationship for error signal input to an integrating gyro torque motor and output for a signal generator*

The governing equation for a gyro device is:

$$T = I\omega\, \partial A/\partial t \qquad (13\text{-}21)$$

where:

T = torque
I = moment of inertia about the spin axis
ω = angular velocity of spin
$\partial A/\partial t$ = angular velocity of precession

The input to the gyro is torque caused by the error signal being fed to small coils about the torque axis shaft. When a torque is applied, the response of the gyro is precession, and a signal proportional to the amount of precession is produced by a signal generator mounted on the precession axis shaft. The ability to follow an input signal accurately is governed by the rate of precession.

$$\partial A/\partial t = \frac{T}{I\omega} \qquad (13\text{-}22)$$

Equation 13-22 illustrates that the rate of precession, and ultimately the response of the gyro, is essentially governed by the spin velocity. The slower the spin, the greater will be the response to a given input torque. However, to achieve a smoothing function, the rate of spin, ω, should be relatively high to prevent the gyro from reacting to all the small changes in the input error signal.

The stabilization process results from the fact that the gyro is mounted in the housing of the tracking element. When the weapon platform pitches and rolls, the gyro reacts to this movement and generates a signal that results in moving the tracking element in the opposite direction to the rotational motion of the weapon platform.

Some limitations exist and must be taken into account when spinning gyros are employed. A gyro spinning on earth will appear to tilt with the passage of time without any force applied. In reality, the gyro is remaining in its original orientation, and the tilt is due to the rotation of the earth on its axis. Because the gyro does not rotate on frictionless bearings, there is additionally a reaction force that also causes the gyro to drift.

Ring Laser Gyro

The ring laser gyro (RLG) uses the same principle of inertial rigidity as a mechanical gyro uses to measure angular velocities and/or accelerations to compute angular displacements; see figure 13-24. RLGs have the distinct advantage of no moving parts. Aside from the obvious reliability advantages, eliminating the friction of internal moving parts means that gyro drift can be reduced to an insignificant amount.



FIGURE 13-24 *Ring laser gyro*[11]

The principal components of an RLG are a laser source and the cavity. The cavity is an enclosed space, usually having the shape of a square or triangle, in which the laser propagates from one corner to the next. Highly efficient mirrors are mounted in the apex of each corner to form a closed "ring." Such a cavity is called a ring laser.

When a laser beam is split in two and propagated in opposite directions around a stationary ring, their respective path lengths are equal, as is the frequency of the laser light, and each beam forms a standing wave inside the cavity. But when the ring is rotated about an axis normal to the ring plane, their respective path lengths and frequencies change.

First demonstrated in 1911, the Sagnac effect notes that two halves of a split beam traveling in opposite directions inside a rotating ring cavity will complete the circuit in different times. This observation has been proved to hold in both relativistic and classical models, and is incorrectly confused with the Doppler effect.[iii] It is the incredible precision of interferometric techniques that makes detecting and measuring extremely small amounts of absolute rotation possible.

The time difference can be correlated to a phase shift of the laser light. Phase differences between the two halves of the traveling laser beam result in a net phase difference, which can be represented by a signal varying in amplitude. Therefore the resulting amplitude is a measurement of phase shift and, consequently, rotation rate.[iv]

The derivation of the expression discussed is as follows:

The amplitude of the output signal for two equal inputs (perfect beam splitter) is:

$$V_{out} = V_{in} \cos\frac{\Delta\varphi}{2} \tag{13-23}$$

where:

V_{out} = voltage at the output detector
V_{in} = voltage supplied
$\Delta\varphi$ = phase difference of the beams upon recombination (in radians)

If the two paths are different by Δx, then the corresponding phase difference is

$$\Delta\varphi = \frac{2\pi\Delta x}{\lambda} \tag{13-24}$$

where:

$\Delta\varphi$ = phase difference
Δx = path length difference
λ = wavelength

To determine the path length difference due to rotation, assume the paths are circular, with radius = R. In absence of rotation, the total path is then $2\pi R$ for each beam. If the entire apparatus rotates at a constant rate (in the clockwise direction) given by ω (radians per second), the two beams will travel different path lengths. The clockwise beam (beam 1) will have to travel an additional amount (depending on time).

$$x_1 = 2\pi R + \omega R t_1 \tag{13-25}$$

where:

x_1 = total distance traveled by beam 1
ω = rotation rate
R = radius of the gyro
t_1 = total travel time for path 1

Since the beam travels at the speed of light (regardless of rotation rate; this is a principle of special relativity), the total distance traveled by the beam in that time is ct_1. In other words:

$$ct_1 = 2\pi R + \omega R t_1 \tag{13-26}$$

Solving for the time:

$$t_1 = \frac{2\pi R}{c - \omega R} \tag{13-27}$$

This corresponds to the intuition that it will take longer than if there was no rotation. The time, then, for the counterclockwise beam (beam 2) to travel the path by a similar argument is:

$$t_2 = \frac{2\pi R}{c + \omega R} \tag{13-28}$$

Note that the sign has changed in the denominator; thus, this beam will take less time than if there were no rotation. The difference in path lengths is then:

$$\begin{aligned} \Delta x &= c\Delta t = c(t_1 - t_2) \\ &= 2\pi cR\left(\frac{1}{c-\omega R} - \frac{1}{c+\omega R}\right) \\ &= 2\pi cR\left(\frac{2\omega R}{c^2 - \omega^2 R^2}\right) \\ &= 4\pi cR^2\omega\left(\frac{1}{c^2 - \omega^2 R^2}\right) \end{aligned} \tag{13-29}$$

But, $\omega R < c$ (otherwise, the outer part of the ring would be traveling near the speed of light!); therefore, we can neglect the contribution of ωR to the denominator:

$$\Delta x \cong \frac{4\pi R^2 \omega}{c} \tag{13-30}$$

Because of rotation, the beams take different times to travel the path length. There will be a net phase shift when they are recombined. This phase shift is given by:

$$\Delta\phi \cong \frac{2\pi}{\lambda}\left(\frac{4\pi R^2 \omega}{c}\right) = \frac{8\pi^2 R^2 \omega}{c\lambda} \tag{13-31}$$

And the resulting effect on V_{out} is:

$$V_{out} = V_{in} \cos\left(\frac{4\pi^2 R^2 \omega}{c\lambda}\right) \tag{13-32}$$

Since all the terms are constants, the output depends only on rotation rate. The RLG therefore measures rotation rate about its sensitive axis. The rotation rate is then found by solving for ω.[v]

RLGs were first constructed in 1963. Since then, fiber-optic cavities and semiconductor laser light, along with low-cost telecom application specific integrated circuits and digital signal processor technologies, have made RLGs compact, lightweight, affordable, and very rugged devices.

Coordinate Systems

To obtain a complete stabilized fire control solution, usually three processes must be accomplished: conversion, transformation, and translation.

A system of coordinates must be established to uniquely describe a position in space or the magnitude and direction of target velocity with respect to a specified reference frame. For location to have meaning, a starting point or a reference point and a system of measure such as length, time, weight, and so on, must first be established.

In order to give the measuring system a zero point and to describe more easily the coordinate system, a reference frame must also be established, as previously mentioned, consisting of three mutually perpendicular lines.

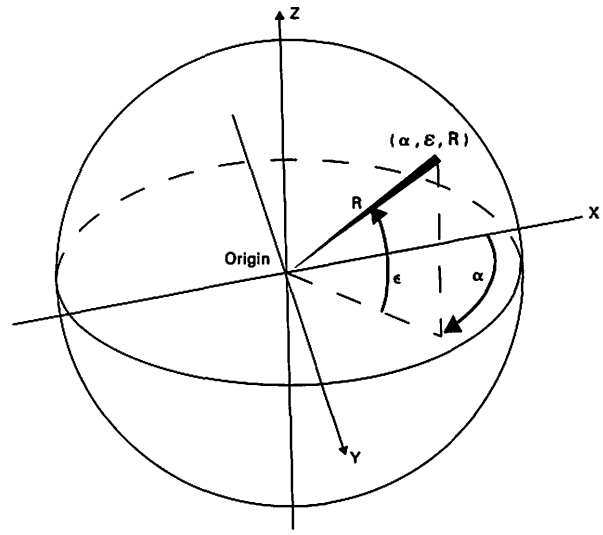


FIGURE 13-25 *Generalized spherical coordinate system*

The point from which these lines originate is the starting point, a zero position.

In actual weapon systems, this zero position is generally the pivotal axes of the sensor and tracking subsystem since all initial target data are generated there. A coordinate, then, is one of a set of values used to locate a point relative to a reference system of lines or surfaces; see figure 13-25.

Any weapon system can be subdivided by coordinate system function into three general categories: data measurement, computation, and data utilization. Because of the nature of hardware design constraints, each functional category employs a type of coordinate system that allows the simplest expression of the problem and the easiest method of actual implementation.

Spherical Coordinate System

A spherical coordinate system is ideal for use in the process of actually measuring target parameters. This is because sensor and tracking systems usually rotate about their coordinate axes and measure distance along a straight line to the target.

A generalized spherical coordinate system consists of two angles and a linear distance, which can be represented by an ordered triplet of characters (α, ε, R) that describe any point in the reference frame; see figure 13-25. Specific examples of the use of the spherical coordinate systems on two diverse real-world weapon platforms are illustrated in figure 13-26.

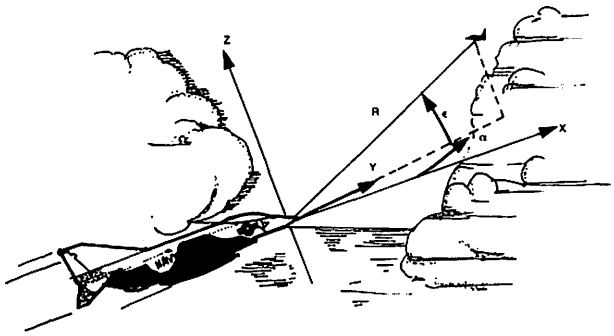


FIGURE 13-26a *Fighter aircraft employment of a spherical coordinate system*

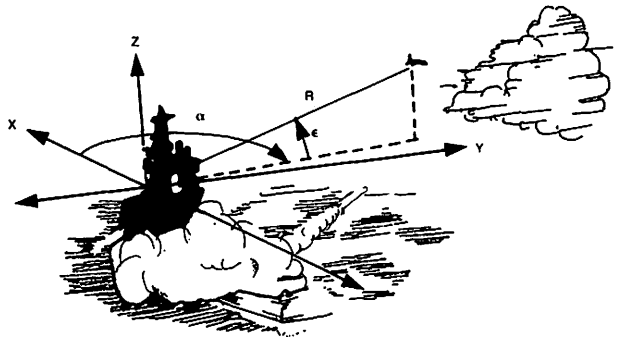


FIGURE 13-26b *Surface combatant employment of a spherical coordinate system*

Cylindrical Coordinate System

Most weapon control computation is accomplished using a rectangular coordinate system, although some older gun fire control systems remain that use a cylindrical coordinate system. A cylindrical coordinate system employs an angle of bearing, a vertical linear distance, and a horizontal linear distance; see figure 13-27.

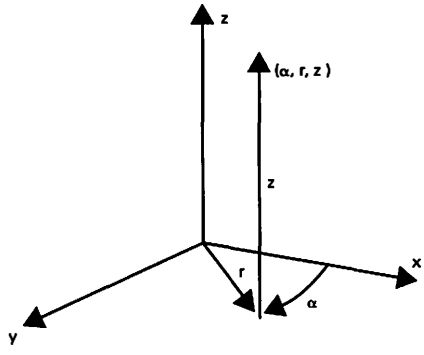


FIGURE 13-27 *Cylindrical coordinate system*

Rectangular Coordinate System

The rectangular or Cartesian coordinate system employs the target's range in the true north-south direction as the *Y* axis distance, the target's range in the east-west direction as the *X* axis distance, and the vertical height as the *Z* axis distance. The rectangular coordinate system is illustrated in figure 13-28.

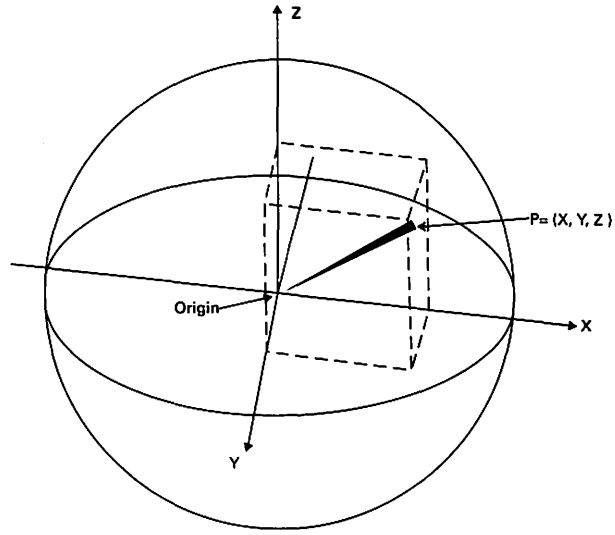


FIGURE 13-28 *Generalized rectangular coordinate system*

Calculation of target velocity and prediction of future target position is easier in Cartesian coordinates, thus reducing the time required for computer solution of the fire control problem. For this reason, as well as the ease with which data can be shared by widely separated forces, the Cartesian coordinate system is used by most modern fire control systems and command and control facilities.

Coordinate Conversion

Coordinate conversion is the process of changing from one system of coordinates that describe a point within a reference frame to another system of coordinates that describe the same point in the same reference frame. Figure 13-29 illustrates the relationship between the spherical coordinate system and the rectangular coordinate system. The equations that govern the conversion are as follows:

$$x = R \cos \varepsilon \cos \alpha \qquad (13\text{-}33)$$

$$y = R \cos \varepsilon \sin \alpha \qquad (13\text{-}34)$$

$$z = R \sin \varepsilon \qquad (13\text{-}35)$$

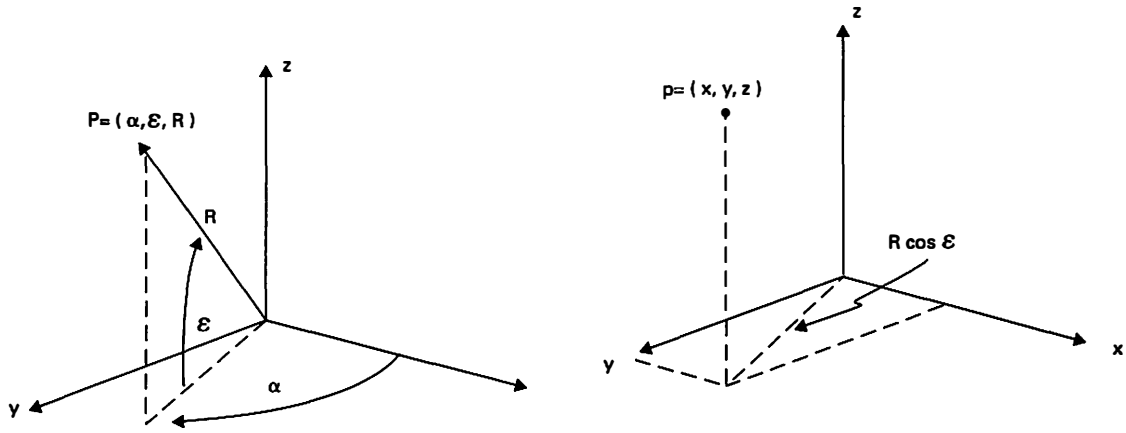


FIGURE 13-29 *The relationship between spherical and rectangular coordinates within the same reference frame*

Coordinate Transformation (Reference Frame Rotation)

The next step in the process of stabilizing the target coordinates is transforming target coordinates from one reference frame to another using the same coordinate system. Because reference frame rotation results from the roll and pitch of the weapon station, the process of transformation involves rotation of the coordinates in the unstabilized frame through the angles of pitch and roll to obtain coordinates in a stabilized frame. Figure 13-30 provides a synopsis of coordinate conversion and transformation.

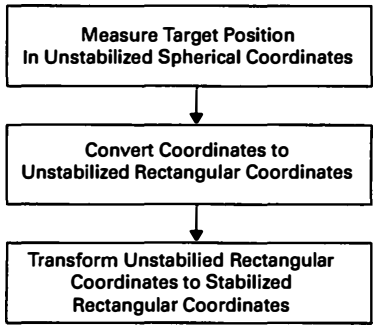


FIGURE 13-30 *Coordinate conversion and transformation*

Consider first a simplified two-dimensional case where the unstabilized coordinates (R, ε) and angle of pitch, ϕ, have been measured. The task is first to convert the measured coordinates into unstabilized rectangular coordinates and then to rotate the unstabilized reference frame through the angle of pitch to obtain rectangular coordinates in a stabilized reference frame.

The basic relationships governing this transformation are:

$$x' = R\cos\varepsilon' \tag{13-36}$$

$$z' = R\sin\varepsilon' \tag{13-37}$$

$$x = R\cos(\varepsilon' + \phi) \tag{13-38}$$

$$z = R\sin(\varepsilon' + \phi) \tag{13-39}$$

Using the formulas for cosine and sine of the sum of the two angles:

$$\begin{aligned} x &= R\cos(\varepsilon'+\phi) = R(\cos\varepsilon'\cos\phi - \sin\varepsilon'\sin\phi) \\ &= (R\cos\varepsilon')\cos\phi - (R\sin\varepsilon')\sin\phi \\ &= x'\cos\phi - \varepsilon'\sin\phi \end{aligned} \tag{13-40}$$

$$\begin{aligned} z &= R\sin(z'+\phi) = R(\sin\phi\cos\varepsilon' + \cos\phi\sin\varepsilon') \\ &= (R\cos\varepsilon')\sin\phi - (R\sin\varepsilon')\cos\phi \\ &= x'\sin\phi - \varepsilon'\cos\phi \end{aligned} \tag{13-41}$$

EXAMPLE 13-2: Two-Dimensional Case

Target Parameters:

$$R = 20{,}000 \text{ meters}$$

$$\varepsilon' = 35° \text{ (0.6108 radian)}$$

Angle of Pitch:

$$\phi = +8° \text{ (0.1396 radian)}$$

1. Convert unstable spherical coordinates to unstable rectangular coordinates:

$$\begin{aligned} x' &= R\cos\varepsilon' \\ &= 20{,}000\cos(35°) \\ &= 16{,}383.04 \text{ meters} \end{aligned}$$

$$\begin{aligned} z' &= R\sin\varepsilon' \\ &= 20{,}000\sin(35°) \\ &= 11{,}471.53 \text{ meters} \end{aligned}$$

Note: A starboard roll and a down-pitch angle will both be considered negative angles.

2. Rotate the unstable coordinates through the angle of pitch using equations (13-40) and (13-41):

$$x = x'\cos\phi - \varepsilon'\sin\phi$$
$$= 16{,}383.04\cos(8°) - 11{,}471.53\sin(8°)$$
$$= 14{,}627.07 \text{ meters}$$
$$z = x'\sin\phi + \varepsilon'\cos\phi$$
$$= 16{,}383.04\sin(8°) + 11{,}471.53\cos(8°)$$
$$= 13{,}639.96 \text{ meters}$$

Equations 13-40 and 13-41 represent the rectangular coordinates in the stabilized reference frame in terms of the coordinates in the unstabilized reference frame. These equations are general in nature and serve to illustrate a fundamental concept in analytic geometry, which can be stated as follows:

> Rectangular coordinates in a reference frame that is rotated with respect to a second reference frame can be expressed in terms of the rectangular coordinates of the second reference frame and the sine and cosine functions of the angle of rotation.

For our purposes, angles of rotation are the angles of pitch and roll of the weapon station. Once rectangular coordinates have been generated for the unstable system, direct applications of the above principle and the general forms of equations 13-40 and 13-41 for angles of pitch and roll will yield rectangular coordinates in the stable system. Figure 13-31 illustrates the relationship between the stable and unstable coordinates when only angle of pitch is considered.

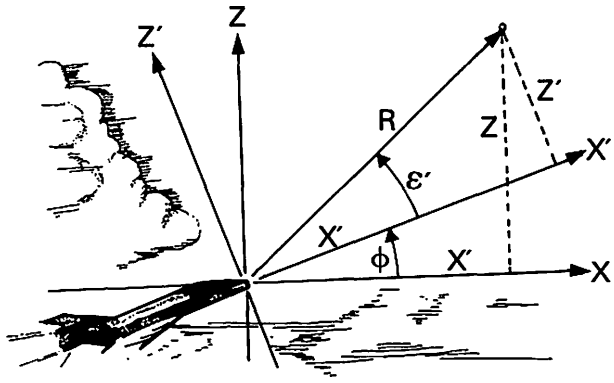


FIGURE 13-31 *Two-dimensional relationship between the unstabilized reference frame (X', Z') and the stabilized reference frame (X, Z), considering only the angle of pitch (ϕ)*

In the general case where roll is nonzero, the coordinates (*X*, *Y*, *Z*) derived above would be an intermediate set of coordinates and would be used to rotate about the *X* axis or the roll axis as follows:

$$\theta = \text{roll angle}$$
$$X \text{ stable} = X \text{ (rotation is about axis of roll)} \quad (13\text{-}42)$$
$$Y \text{ stable} = Y\cos(\theta) - Z\sin(\theta) \quad (13\text{-}43)$$
$$Z \text{ stable} = Y\sin(\theta) + Z\cos(\theta) \quad (13\text{-}44)$$

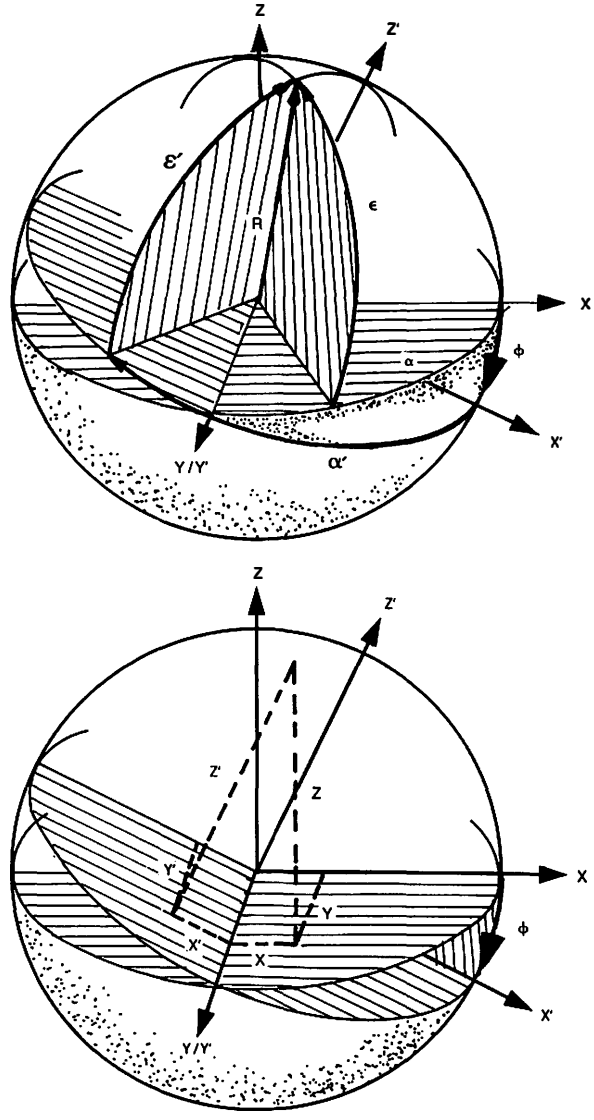


FIGURE 13-32 *Relationship between reference frames and coordinate systems when azimuth is not zero and rotation is through the angle of pitch (ϕ)*

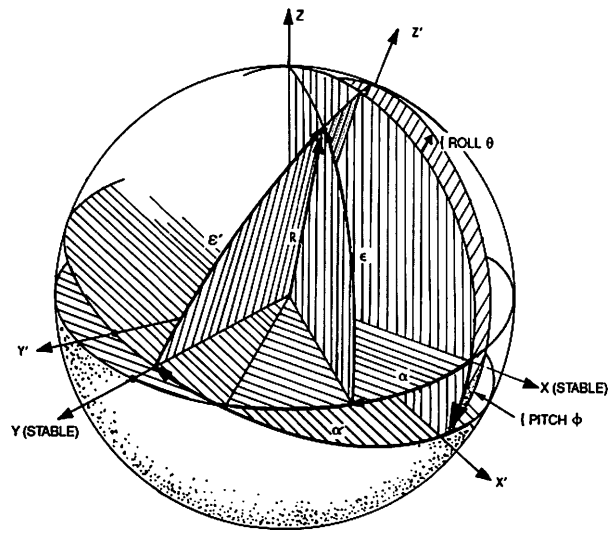


FIGURE 13-33 *Spherical coordinate relationships that consider both pitch and roll*

The general scheme for reference frame rotation is as follows:

Convert unstable spherical coordinates to unstable rectangular coordinates.

$$x' = R\cos\varepsilon'\cos\alpha' \tag{13-45}$$

$$y' = R\cos\varepsilon'\sin\alpha' \tag{13-46}$$

$$z' = R\sin\varepsilon' \tag{13-47}$$

Rotate through the angle of pitch:

$$X = X'\cos\phi - Z'\sin\phi = (R\cos\varepsilon'\cos\alpha')\cos\phi - (R\sin\varepsilon')\sin\phi \tag{13-48}$$

$$Y = Y \text{ (rotation is about the } y' \text{ axis)} \tag{13-49}$$

$$z = x'\sin\phi + z'\cos\phi = (R\cos\varepsilon'\cos\alpha')\sin\phi + (R\sin\varepsilon')\cos\phi \tag{13-50}$$

Rotate through the angle of roll:

$$X \text{ stable} = X \text{ (rotation is about axis of roll)} \tag{13-51}$$

$$Y \text{ stable} = Y\cos\theta - Z\sin\theta \tag{13-52}$$

$$Z \text{ stable} = Y\sin\theta + Z\cos\theta \tag{13-53}$$

Reference Frame Translation

Reference frame translation can be defined as "the location of a point (target) with respect to two different reference frames whose origins are physically displaced from one another." Figure 13-34 illustrates this relationship for two coplanar reference frames, *A* and *B*, that are displaced linearly by distance *h*.

The point *T* (target) may be located in direction and distance from reference frame *A* by the angle ϕ_1 and the range R_1. In order to transform target coordinates from reference frame *A* to reference frame *B*, the angle P_h must be taken into consideration to provide the coordinate ϕ_2 and R_2.

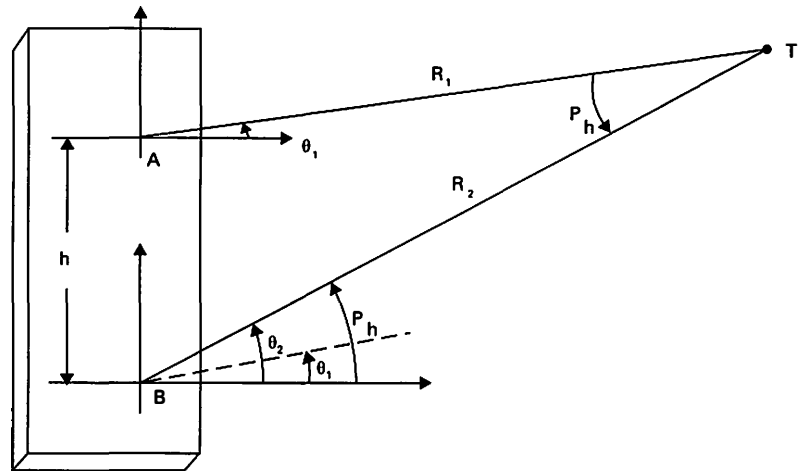


FIGURE 13-34 *Effect of translation on target location*

This displacement of reference frames is common aboard ships. It is necessary, for example, to describe a target's location accurately in both the reference frame of a missile director and the frame of its associated missile launcher. The term *parallax correction* is used to describe the data conversion needed to compensate for this translation of reference frames.

In addition to the horizontal displacement, illustrated in figure 13-34, a vertical displacement also exists due to the general configuration of directors and launchers (or guns) being mounted on different deck levels; see figure 13-35.

Weapon System Alignment

Modern weapon systems ashore and afloat include many elements that are physically displaced from one another. To function as an effective system, these elements must be aligned to one another in bearing and elevation and must rotate in parallel planes. When this is accomplished, the following requirements are met:

1. All weapon bores, launcher rails, sight telescopes, and radar beams are parallel and remain parallel throughout any operating motion (when no parallax or ballistic corrections are made).
2. All related readouts of movements are correct with respect to the established reference.
3. All intra-element angle transmissions are correct.

To meet these requirements, the following specifics must be accomplished:

1. All readouts and transmission systems are aligned so that the 0° train reference lines of the elements are parallel.
2. The planes of rotation (the roller-path planes) of the elements are parallel or effectively parallel.
3. All readouts and transmission systems are aligned so that when the pointing lines are parallel in train, the zero elevation references are parallel.

Alignment

Alignment of weapon system elements begins with establishment of the parallel planes of rotation of the system elements. Ashore, in Marine systems, surveyors and combat engineers will determine the suitability of the chosen site with reference to the slope of the ground, the physical space available, and the presence of obstructions.

Truck- or trailer-mounted elements must be firmly emplaced and then leveled using adjustable braces provided for that purpose. Once this has been done, the only error would result from defects in the machined surface of the truck or trailer frame upon which the element rotates.

Because a ship changes shape when afloat (due to loading) and is in constant motion, the initial phase of the alignment of launchers, sensors, and navigational references is performed in dry dock. Once parallel roller-path planes are established with the ship constrained by keel blocks, the ship is floated, loaded to a specified displacement, and the data rechecked. Sometimes machining of the roller path is required afloat due to the ship's bending or sagging when waterborne.

Whether aboard ship or ashore, one element is selected as a reference, and all others are aligned to it. Aboard ship, a forward gun director or, in the absence of a gun director, the lowest forward missile director is selected. Ashore, the element with the greatest precision is made the reference. In any case, it is helpful for the reference element to be in visual sight of the other elements for train and elevation alignment.

Once parallel planes of rotation have been established, the 0° train and elevation references must be established for each element, and the display of angular measurements must be calibrated. Transmission of this data between elements must be verified for accuracy, and a correction applied if necessary.

Ships or shore-based elements will employ calibrated metal bars or yokes, called tram bars, as gauges to verify the accuracy of angular measurement in periodic rechecks after alignment has been performed. The tram bar is placed between metal pads, one of which is located on the stationary base and the other on the rotating part of the element. This establishes a calibrated angle of rotation



FIGURE 13-35 *Horizontal and vertical parallax considerations for fire control radar to gun systems*

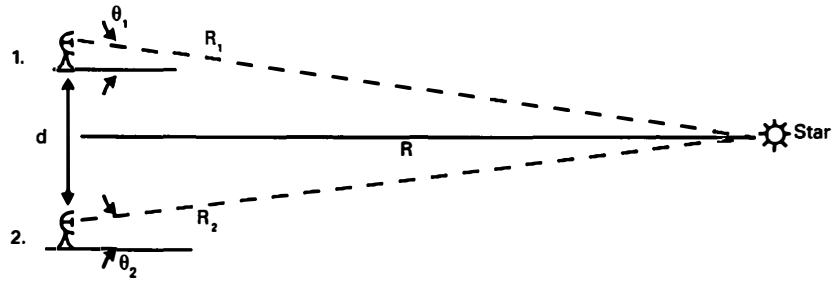


FIGURE 13-36 *As* R *goes to infinity,* d *becomes much less significant; therefore, the angles measured by* θ_1 *and* θ_2 *will approach equality.*

that is compared with readouts of the transmission systems; in the case of an error, they are adjusted to read the calibrated angle that should correct the error throughout the range of the element's motion.

One of the best ways to verify alignment between elements is by performing the "star check." In this procedure, a star is selected of proper elevation at a convenient bearing, and all system elements are trained and elevated to it. Launchers and guns will have temporary optical boresights installed to perform the check. Stars are such a great distance away that the parallax between elements is essentially zero; therefore, all elements should have the same measurement of train and elevation when sighted on the same star; see figure 13-36.

Collimation

Many system elements such as missile fire control radars may employ several different radar beams plus an optical line-of-sight (LOS) reference. The parallelism between beams radiated by a single element is achieved by the process of collimation. At a specified distance, the element is sighted upon an optical target that has radar-receiving antennas displaced from it.

When each element is properly aligned, the center of the beam will coincide with the optical target; see figure 13-37. The element is trained and elevated back and forth across the target while power measurements are taken. If maximum signal strength is measured at some time other than when the optical target is in the center of the boresight telescope, a mechanical or electrical adjustment is made to the antenna to correct it.

Aircraft Boresighting

The boresighting process accomplishes the same function for aircraft as collimation does for ships and land-based systems. Its purpose is to establish a relationship between

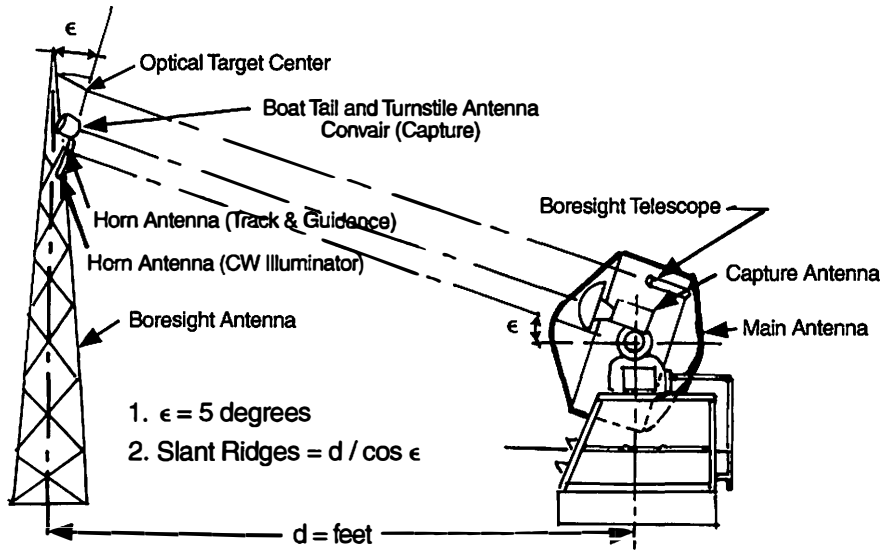


FIGURE 13-37 *Collimation of a multiple-beam fire control radar*

flight attitude of the aircraft, bore axes of guns and weapon-releasing equipment, radar beam axes, and those of optical sights and head-up displays (HUD).

The process is designed to employ pertinent ballistics and mechanical data in establishing the relationship between the sight line of the optical or radio frequency aiming device and the trajectory of the weapon being fired or ejected, in order to predict its mean point of impact at predetermined ranges.

The boresighting procedure is carried out in two phases: electrical boresighting and harmonization. Electrical boresighting requires that RF energy be radiated, but harmonization can be accomplished without radiation if the antenna's optical axis is in proper agreement with its RF axis.

The electrical boresighting procedure is the equivalent of collimation as presented previously. The aircraft's installed radar or radars are made parallel to the optical axis of the antenna. A portable frame is attached to the aircraft—see figure 13-38—that incorporates radar and optical targets for various armament and sensors. This frame is employed much like the collimation tower in figure 13-37.

The harmonization procedure brings the radar optical axis into a parallel relationship with a baseline in the aircraft structure called the armament datum line. This is the

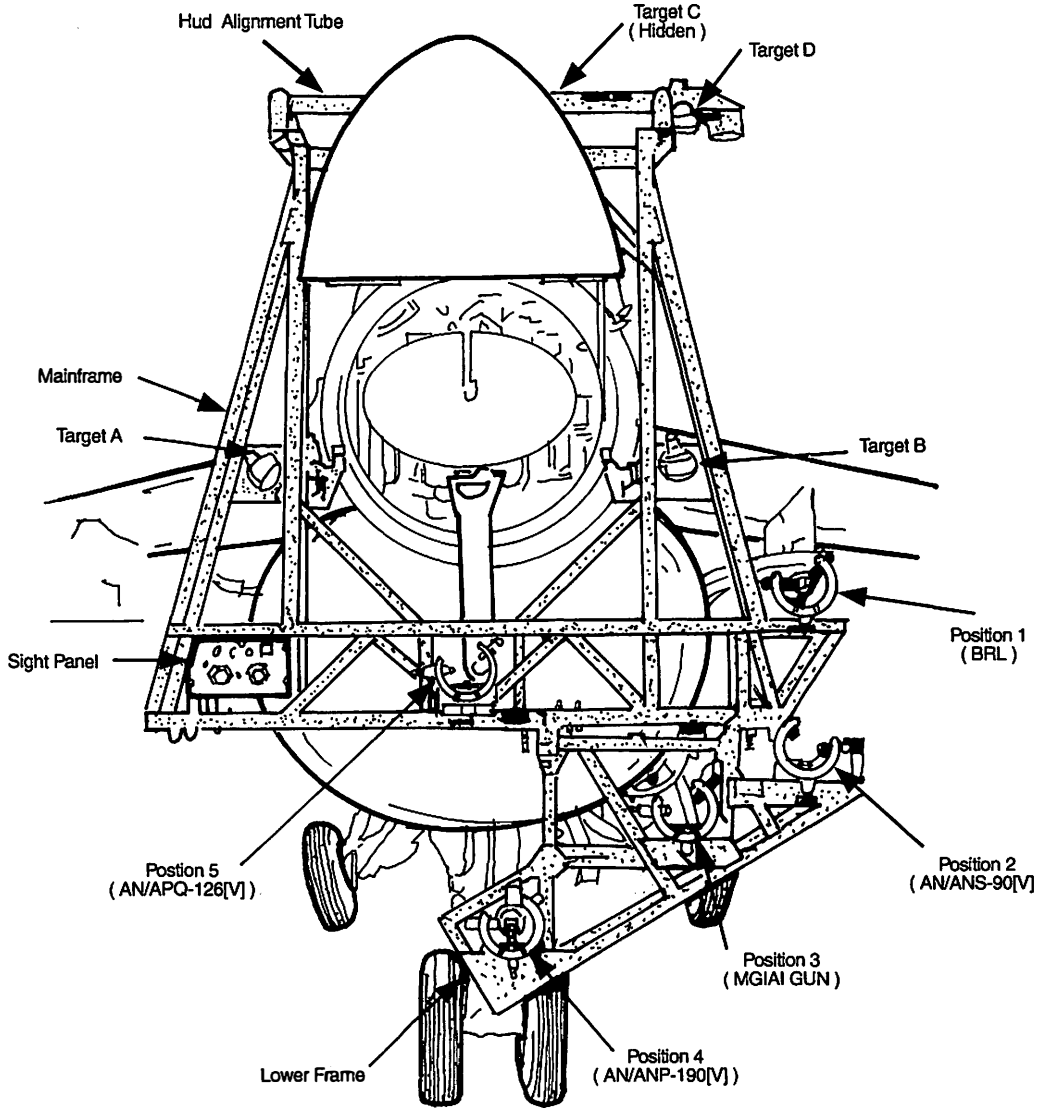


FIGURE 13-38 *Short-range boresighting equipment*

equivalent of the process of roller-path plane of rotation adjustment for ships in addition to the establishment of zero train and elevation. Harmonization also includes alignment of the weapon-releasing equipment and optical sights or HUD.

Summary

This chapter has described general procedures for the solution of a weapon control problem. The overall prediction procedure was divided into subprocedures of relative motion, weapon ballistics, and transferring between stabilized and unstabilized inertial reference frames.

As input, the relative motion procedure receives present target position, relative target velocity, and weapon time of flight and yields as an output successive approximations to future target position.

A ballistics procedure receives as input the approximations to future target position, and its output is successive approximations to the launch angle and time of flight.

The overall iterative process is terminated when $T_n - T_{n-1} \rightarrow 0$ (figure 13-11), which indicates that the refinement of future target position is becoming precise.

Figure 13-39 provides a summary of the weapon control problem in flowchart form.

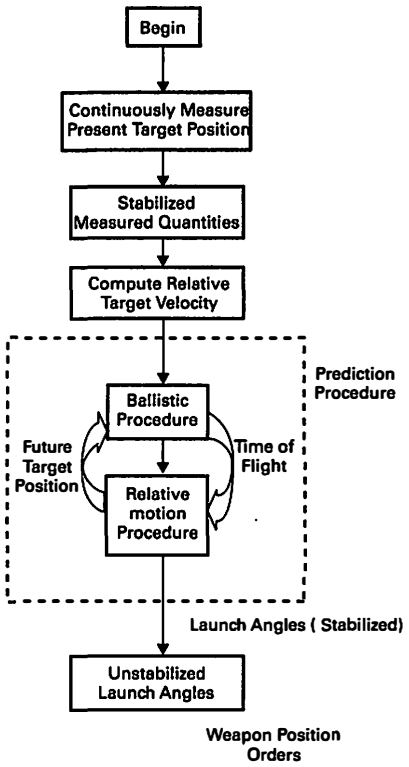


FIGURE 13-39 *Summary of the solution to the fire control problem*

---

[i] Corse, 54

[ii] Courtesy of Honeywell

[iii] http://www.mathpages.com/rr/s2-07/2-07.htm

[iv] *FAS*

[v] *FAS*

14 Weapon Propulsion and Architecture

Introduction

Every weapon requires some type of propulsion to deliver its warhead to the intended target. This chapter is a study of the propulsion and launching systems used to get the weapon off the firing platform and to propel weapons to their targets. Design requirements for weapon-type vehicles are environmentally dependent and require knowledge of the forces affecting the weapon in motion.

The underlying principle of propulsive movement was stated by Isaac Newton in his Third Law of Motion: For every action there is an equal and opposite reaction. Every forward acceleration or change in motion is a result of a reactive force acting in the opposing direction—for example, a person walks forward by pushing backward against the ground.

In a propeller-type airplane, the air through which it is moving is driven backward to move the airplane forward. In a jet-propelled plane or a rocket, a mass of gas is emitted rearward at high speed, and the forward motion of the plane is a reaction to the motion of the gas.

Matter in liquid, gas, or solid form may be discharged as a propellant force, expending its energy in a direction opposite to the desired path of motion, resulting in a predetermined acceleration of the propelled body along a desired trajectory.

Launching System

The purpose of a launching system is to place a weapon into a flight path as rapidly as possible. Flight initiation occurs as a result of the propelling forces of gravity, impulse, or reaction being imparted to the weapon. Launching must occur at an optimum moment to ensure that the weapon system functions effectively. The launching system must also safely withstand the propulsive forces of the weapon, and must be highly reliable in order for the weapon to achieve the damage probability for which it was designed.

The most significant requirements of launching systems are:

1. *Speed*. The launcher must be capable of rapid initial employment and a subsequent high rate of fire.
2. *Reliability*. No matter what the degree of sophistication adopted in the design of a launcher, it must be capable of a certain level of repeated use without failure. It must also be repairable.
3. *Safety*. The vehicle upon which the launcher is mounted and the personnel who operate and control the launcher must be able to function without damage or injury.
4. *Compatibility*. The launching system must complement the mission of the delivery vehicle and, consequently, other weapon systems installed. It must be designed to withstand the particular rigors of deployment associated with the mission of the delivery vehicle in a potentially corrosive environment with strong dynamic forces and vibrations.

Impulse Launchers

An impulse launch occurs when force is applied to the weapon either to project it along the entire path of its trajectory or, to a lesser degree, to clear the weapon from the delivery vehicle.

Two basic launcher types employ this method: cannon launchers and ejector launchers. Cannon launchers are more common and may be further subdivided into guns, howitzers, and mortars. The basic differences among these subdivisions are trajectory, initial velocity, and size of the launcher, with guns having the highest muzzle velocities and flattest trajectories. Since the functional operation of the three cannon launcher types is similar, gun launchers will be examined in detail.

Gun-Type Launchers

A gun launcher not only imparts initial velocity (IV) to the weapon (more commonly referred to as a projectile), but also provides projectile flight guidance in the most simple application. Additionally, specially designed projectiles may be fitted with a rocket motor or a guidance package—for example a Laser, IR, or GPS, system. A major consideration in gun launcher design is energy dissipation of the recoil force created by the impulse to propel the projectile.

In chapter 13, interior ballistics and the generation of heat and gas in a confined space were discussed. Gun barrels must be constructed with enough inherent strength to withstand the pressures developed by the propellant. Additionally, most gun launchers are characteristically large and heavy since the force of the initial impulse is largely absorbed by the launcher in conjunction with its recoil/counterrecoil system.

A gun barrel is normally thickest at the powder chamber, where the greatest pressure effects occur. The gun barrel then tapers in thickness proportionally to the pressures exerted by the propellant. To gain an appreciation of the basic principles involved in gun design, see figure 14-1.

Gun strength must exceed the powder pressure at every point by an amount that will provide a sufficient margin of safety. The pressure curves in figure 14-1 show pressure beginning at a value well above zero. This indicates that there is a pressure buildup that occurs after the propelling charge begins to burn but before the projectile begins to move.

Note that the gun strength curve does not vary parallel to the powder pressure curve. The reason is that the same pressure that the expanding gases exert against the base of the projectile is exerted equally against all interior surfaces of the gun behind the projectile. Hence, the breech part of the barrel must be designed to handle the maximum stress.

After the projectile passes the point of maximum pressure, it continues to be accelerated by gas pressure until it leaves the muzzle. The total area under the curve, up to the point where the projectile leaves the gun, is a rough measure of initial velocity, and the pressure remaining at the muzzle is an indication of the muzzle loss. Also, a high muzzle pressure increases muzzle flash.

As can be seen in the figure, employing a high explosive would be catastrophic, exceeding the strength of the gun almost instantaneously upon detonation. In fact, the release of energy by an explosive would be so rapid that the greater portion would be expended in rupturing the gun rather than in translating the projectile through the barrel.

Two principal stresses act on the gun barrel during normal propellant burning: a tangential stress or tension coupled with a radial stress (tending to split the gun open longitudinally) and a longitudinal stress (tending to pull the gun apart in the direction of its axis). The greatest stress is in a direction tangent to a radius of the barrel, due to gas pressure, and thus the longitudinal stress can be neglected without significant error.

Lames' law states: "At any point in a cylinder under fluid pressure, the sum of the tangential tension and the radial pressure varies inversely as the square of the radius." This means that in a cylinder under internal pressure, points in the metal close to the bore experience a large proportion of the stress, whereas those at a greater radius experience progressively less stress.

The result is a limit in thickness beyond which any additional wall thickness will not add much in strength. Therefore, construction methods must be employed that will cause the outer layers of metal to absorb more stress.

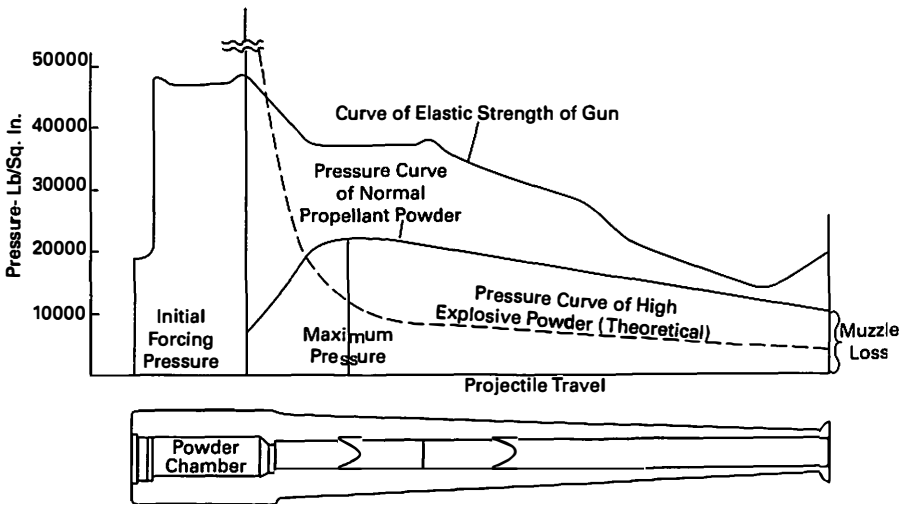


FIGURE 14-1 *The relation between the gun barrel cross section and strength/pressure curves. (The X-axis in the figure represents projectile movement in the bore, not time or bore length.)*

All modern gun barrels are made of steel, and they are generally prestressed to make them more resistant to internal (bursting) pressures. The object of prestressing is to make the outer layers of metal in the barrel bear a greater proportion of the bursting load.

The built-up method of prestressing is to heat steel ring-shaped jackets, or hoops, to high temperatures and then to slip them over the gun tube and allow them to cool. As the hoops cool, they contract, until at the end of the process, they squeeze the tube with a pressure of thousands of pounds per square inch; see figure 14-2.



FIGURE 14-2 *76 mm naval gun; notice the metal jacket over the gun tube at the end of the barrel.*

Another prestressing technique applied to modern barrels of one-piece or monobloc construction involves radial expansion. In this process, a gun tube with a bore slightly smaller than the caliber desired is expanded by hydraulic pressure. When the pressure is released, the outer layers of the tube tend to return to their original dimensions, whereas the enlarged inner layers (liner) tend to maintain their enlargement.

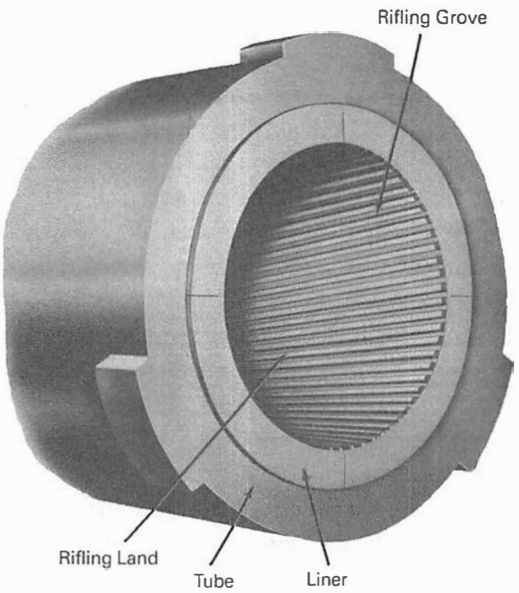


FIGURE 14-3 *End of gun barrel showing liner-tube arrangement*

Thus, the inner layers of metal are severely compressed by the contraction of the outer layers, as if a hoop had been shrunk onto them. The built-up and radially expanded methods may be incorporated in a single gun. The 5"/62-caliber gun, for example, has a jacket shrunk onto a radially expanded tube; see figure 14-3.

Smaller guns are made from a single steel forging with neither radial expansion nor hoops. The pressure per unit area may be higher in small guns than in large guns; however, the tube walls may be made thicker and more massive in small guns without resulting in an excessively large forging. This type of construction is limited to guns of 3-inch caliber and smaller.

Recoil/Counterrecoil Systems

Recoil is rearward movement of the gun and connected parts during and after firing. It is caused by reaction to the forward motion of the projectile and propellant gases. After recoil, the gun and connected parts return to the in-battery, or firing, position. This forward movement is called counterrecoil.

If the gun were mounted rigidly, without any recoil system, it would be practically impossible to build a carriage to withstand the loads imposed upon it, without rupturing, overturning, or displacing it. To bring the carriage stresses down to a reasonable value and to ensure stability, a recoil system is interposed between the gun and carriage.

The recoil mechanism absorbs energy of the recoiling parts over a certain convenient length and returns the gun to battery for further firing. The recoiling parts are

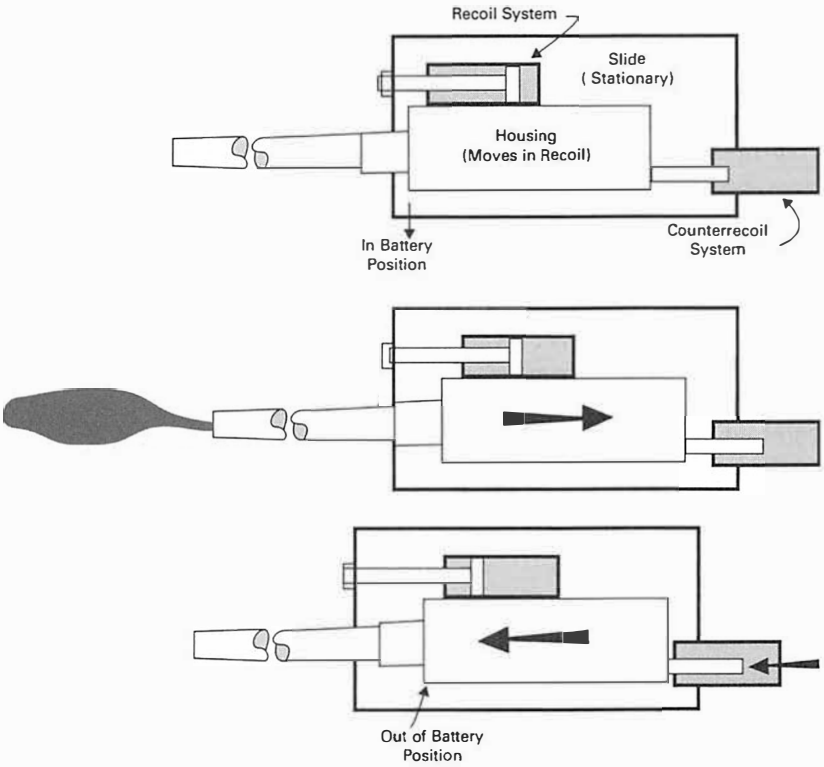


FIGURE 14-4 *Recoil/counterrecoil of a medium-caliber gun*

the gun and those parts of the recoil mechanism and carriage that move with the gun in recoil and counterrecoil.

A recoil mechanism usually consists of three components: a recoil brake, which controls the recoil and limits its length; a counterrecoil mechanism, which returns recoiling parts to the in-battery position; and a counterrecoil buffer, which slows the end of counterrecoil motion and thereby prevents shock to the recoiling parts.

Figure 14-4 illustrates the recoil/counterrecoil sequence when a gun is fired. The top diagram illustrates the main parts of the system. When the gun is fired, thrust of recoil moves the housing rearward—middle diagram. After the recoil system has caused the gun slide to stop its rearward motion, the counterrecoil pressure causes the housing to move forward to an in-battery position—bottom diagram.

In most naval guns, the recoil brake is hydraulic and consists of four elements: a cylinder, a piston, a liquid such as glycerin and water, and some form of orifice in the piston that allows the liquid to flow from one side of the piston to the other. The motion of the piston within the cylinder forces the liquid through the orifices, which absorbs the energy of recoil and controls the return of the gun to in-battery during counterrecoil.

The work required to force a liquid through a given orifice is determined from the laws of hydraulics and depends on area of the orifice, area of the piston, velocity of the piston, and properties of the liquid.

It can be shown that the work absorbed by the hydraulic brake can be fully accounted for in the rise in the temperature of the liquid. Under rapid-fire conditions, the temperature rise is cumulative from shot to shot and results in a considerable rise in temperature, which must be taken into account in designing a recoil system.

Muzzle Brake

Another method of reducing the forces of recoil is the muzzle brake, which is a device consisting of one or more sets of baffles attached to the muzzle end of the gun barrel; see figure 14-5. As the projectile leaves the muzzle,

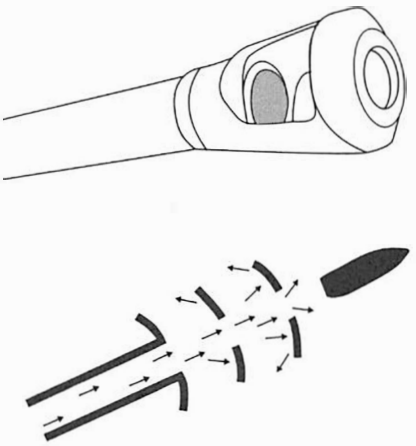


FIGURE 14-5
Muzzle brake

high-velocity gases following the projectile through the tube strike baffles of the brake and are deflected rearward and sideways into the atmosphere.

In striking the baffles, the gases exert a forward force that partially counteracts the force of recoil. The muzzle brake also acts as a blast deflector and reduces obstruction of the target, an important function in direct-fire weapons such as tanks.

Soft Recoil Systems

In the development of gun launchers for aircraft, it was found that the high-velocity projectiles caused excessive forces at the gun trunnions. Consequently, additional structural strength and thus weight had to be built into the airframe to absorb the stresses transmitted to the aircraft structure. To reduce these forces, the principle of soft recoil was applied to these weapons. Another term for this principle is *out-of-battery firing.*

The soft recoil system differs from conventional recoil systems in that the recoiling parts are mechanically held against a gas spring force acting in the direction of firing. Upon release, the expanding gas accelerates the housing forward in the direction of fire; propellant ignition occurs automatically when proper velocity is attained.

The firing impulse overcomes the forward momentum of the housing, reverses its direction, and forces the housing back against the gas spring until latched into the battery position. Figure 14-6 illustrates the difference between the conventional and soft recoil systems.

The advantages of the soft recoil system are that it greatly reduces the horizontal forces on the gun platform or trunnions and that the total recoil cycle time is reduced, allowing for higher rates of fire. In addition to aircraft guns, this system is currently employed on the 105-millimeter light-towed howitzer.

Recoilless Systems

Another solution to the recoil problem is the recoilless principle. The recoilless rifle operates on the basis that if the momentum of the propellant gases discharged to the rear is equal to the gases expelled forward and projected, then the launcher will have no momentum imparted to it. Considering the four movable parts of the system to be the projectile, the gases forward, the gases rearward, and the gun itself; and representing the mass of each as m_1, m_2, m_3, and, m_4, and the velocities of each as v_1, v_2, v_3, and v_4, respectively, then:

$$m_1v_1 + m_2v_2 + m_3v_3 + m_4v_4 = 0 \qquad (14\text{-}1)$$

If the gun is recoilless, then $v_4 = 0$, and the equation reduces to:

$$m_1v_1 + m_2v_2 + m_3v_3 = 0$$

or (14-2)

$$m_3v_3 = -m_1v_1 - m_2v_2$$

That is, the combined momentum of the projectile and the forward-moving gases is equal, but opposite in direction, to the momentum of the gases discharged through the breech. If this is true for all instants during the weapon's firing period, which is about a hundredth of

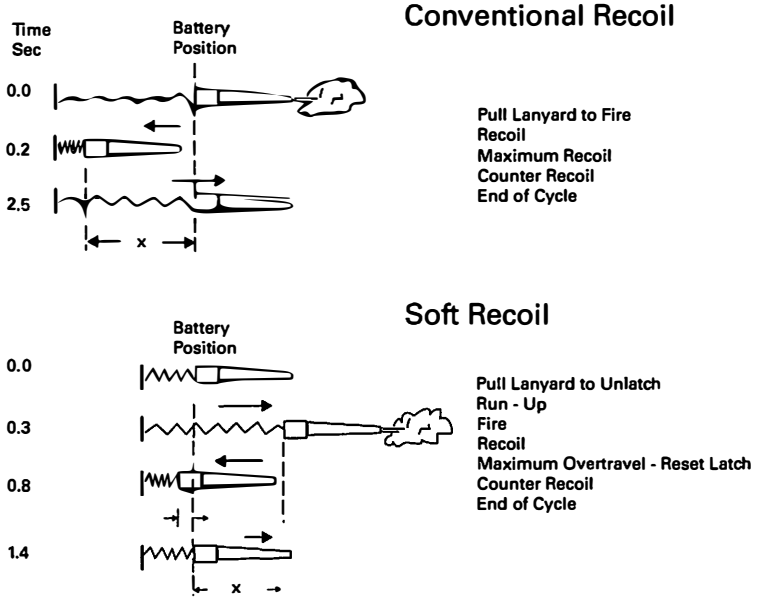


FIGURE 14-6 *Comparison of conventional recoil to soft recoil*

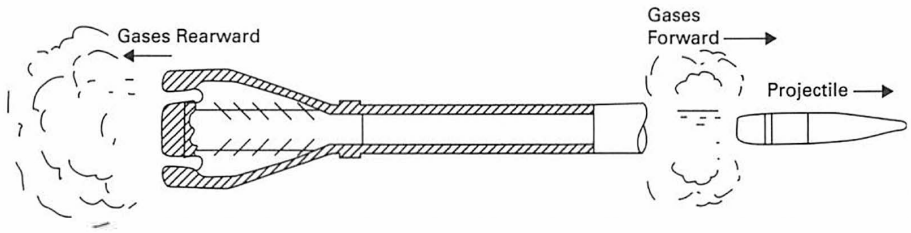


FIGURE 14-7 *Recoilless rifle*

a second, then the momentum of the weapon must equal zero at all times during firing, and the weapon can be considered perfectly recoilless.

In practice, however, this is generally not the case. A recoilless weapon is usually recoilless in the mean—that is, although the total momentum applied to it over the firing period is zero, the sum of the momentums at any instant is not necessarily zero. The weapon undergoes large unbalanced forces during some parts of the pressure interval, and oppositely directed forces during other parts of the pressure interval. Thus, its "recoillessness" is an average rather than an absolute value. Figure 14-7 illustrates the basic considerations associated with a recoilless rifle.

The main advantage of recoilless rifles over conventional guns of comparable caliber is their much lighter weight, which gives them greater mobility. The price paid for this advantage is a slight reduction in velocity and range and a large increase (2.5 to 3 times) in the quantity of propellant required.

The main drawback of recoilless launchers is the tremendous blast to the rear of the gun (back blast) caused by the escaping gas. The danger zone for a comparatively low-caliber (57-millimeter) rifle is a conical section 17 meters long and 14 meters wide diverging rearward from the breech. Because of these disadvantages and the fact that techniques have been developed for achieving greater mobility of conventional guns, recoilless rifles are not as prevalent in the ground arsenal as they were in the past.

Ejector Launchers

Impulse launchers for weapon ejection are employed for both free-fall and self-propelled weapons. Their main purpose is to ensure that the weapon safely clears the delivery vehicle. Ejection is usually accomplished by expansion of high-pressure gases from a compressed air supply or from ignition of a propellant charge. Because it is used for ejection purposes only, the impulse is small, and the launcher can be built to withstand the shock of launching without the need for excessive structural strength or special devices. Thus, launchers of this type are fairly light and simple in design.

Self-propelled weapons, which are impulse launched, frequently are large and heavy, making transfer and loading a slow, laborious process. A launcher of this type usually has several launching tubes, and the weapons are stored within the tubes until needed.

The most common launcher of this type is the torpedo tube. Torpedo tubes are installed in ships and submarines and may be of fixed or trainable design. Torpedo ejection is caused by expanding gas from a gas-generating chemical or from a compressed air supply. In earlier submarines, fixed torpedo tubes were installed in the bow and stern; in present U.S. Navy submarines, however, they are located only in the forward section.

Fixed tubes are also found in certain surface combatants. These nontrainable tubes are mounted within enclosed spaces singly or in groups of two or more on each side of the ship. To preclude the launch vehicle having to maneuver radically to provide ideal orientation for the torpedo relative to the target at launch, torpedoes deployed from fixed tubes are equipped with guidance features to compensate for this restriction. The interior location permits all tubes to be located in a torpedo compartment where they may be readily maintained, regardless of weather or topside activities.

Trainable tubes are often installed aboard surface ships in a clear deck area so that they may be oriented as required. The triple-tube launcher, illustrated in figure 14-8, is most commonly found on destroyer-type combatants.

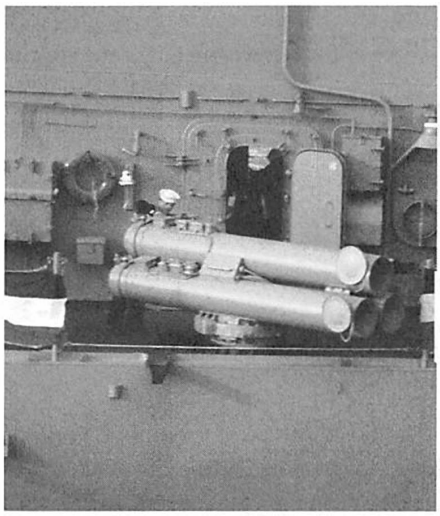


FIGURE 14-8
MK-32 torpedo tubes employed aboard most surface combatants as a launcher for antisubmarine torpedoes

To reduce space, the tubes are stacked one above the other. By fabricating the tubes of reinforced fiberglass, weight is reduced to a minimum, and maintenance due to topside exposure is less. In this figure, note the air flasks at the breech. Removable covers are provided at the muzzle when the launcher is in a stowed configuration.

If compressed gas is not used, ejection may be accomplished by burning a propellant grain. In this case, the hot gases generated by the propellant grain may be used to flash water to steam. The steam/gas mixture can then be introduced to the eject pressure chamber to eject the weapon. This steam-ejection-type system is found aboard ballistic missile submarines. The result is a flameless launch, with rocket motor ignition of the missile occurring at a safe distance from the submarine.

Ejection launchers are also employed aboard aircraft. Nearly all free-fall and glide weapons carried by aircraft are ejection-launched. At the time of launch, high-pressure gas acts on a piston, forcing the hooks holding the weapon to open, and simultaneously forcing a metal rod or ejection foot against the weapon.

This action physically ejects the weapon away from the aircraft, ensuring that it is clear of the aerodynamic flow around the aircraft. This high-pressure gas is usually formed from explosive cartridges, but could also be from a stored compressed-gas cylinder. In addition to free-fall weapons, some missiles are also ejection-launched to clear them from the aircraft prior to motor ignition.

Electromagnetic Impulse

Electromagnetic propulsion is a novel impulse-launching concept. Instead of using gas pressure to move a projectile, this concept uses electromagnetic forces that are generated by the interaction of an electric current and a magnetic field. When a current is at right angles to a magnetic field, a force is produced at right angles to the current and the magnetic fields as predicted by Faraday's law of induction.

$$\vec{F} = i\vec{L} \times \vec{B} \qquad (14\text{-}3)$$

where:

$\vec{F}$ = force vector
i = current
$\vec{L}$ = length of rail vector
$\vec{B}$ = magnetic field vector

One method that uses this type of propulsion is essentially a series of solenoids moving a projectile down a hollow tube, called a coil gun. The coil gun uses successive coils of wire around a tube that interacts with a separate coil wrapped around a projectile as it is drawn down the tube.

This method of movement requires precise timing of the successive tube coils to move the constantly accelerating projectile the length of the tube. Because of the precise timing involved and the need for the projectile to produce a separate voltage to apply to its own coils, this method of launch is not receiving much attention.

A more feasible method for future electromagnetic launching armament applications is called the rail gun. Rail gun construction is simply two parallel highly conductive rails with grooves cut in and fitted with a projectile. The projectile has a conductive armature usually made of aluminum, copper, or graphite that fits into the grooves of the rails to complete a circuit.

Rails are connected to a large power source that supplies current to one rail. The current then passes through the armature and back through the other rail to the power source, thus completing the circuit. This electric current interacts with the magnetic fields of the rails and moves the projectile down the rails at an ever-increasing velocity; see figure 14-9. Since the force is proportional to the square of the current that passes through the rails, if a large enough current were used, the projectile should leave the gun with an enormous velocity. Figure 14-10 shows a preliminary design of a rail gun launch package.

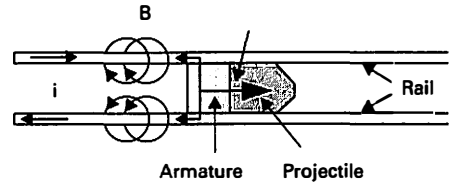


FIGURE 14-9 *Typical rail gun design*

To make rail gun launching systems suitable for weapon employment, a number of problems must be solved. The power supply must be able to generate a large current nearly instantaneously. Laboratory methods of using large battery banks or capacitors have proved difficult to scale for useful military applications. Thus, compact power supplies using rotating machines to store and also generate the current pulse are being developed.

Another problem is wear on the rails themselves. With the large amount of current necessary to move the projectiles at desired velocities of over 2,500 m/s, the rail material wears away rapidly, causing subsequent rounds to have poor contact with the rails and reducing current flow and efficiency.

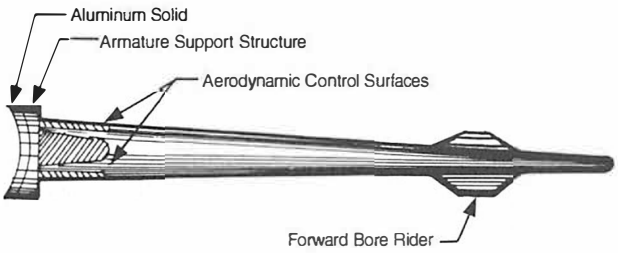


FIGURE 14-10 *Preliminary electromagnetic rail gun launch package concept for naval fires support application*

Still another concern is the large amount of heat generated by electrical resistance of the rails; therefore system cooling is extremely important so the rails do not bend. Much research still needs to be done in all these areas in order to achieve a working weapon design and to replace the chemical-type gun system.

Reaction Launchers

This general type of launcher uses the force contained within the weapon to separate the weapon from the launcher. These weapons are normally rockets or missiles. The propulsion system of the missile itself may be used to provide the necessary force. Thus, most self-propelled weapons, if not launched by ejection, are put into flight by their own reactive forces propelling them off the launchers.

Reaction launchers provide static support for the weapon and initial flight orientation. They are characteristically small and light since they are not required to sustain large moments of force upon weapon launch. Reaction-propelled air weapons often depend upon wings or fins to provide lift, and must use rocket thrust to overcome gravity temporarily and be propelled to the desired flight speed.

If, during the development of thrust, a weapon is free to move along the launcher, it might not have sufficient thrust or lift to overcome gravity by the time it leaves the launcher. Thus, the missile could fall to the deck of a launching ship or become completely disoriented before sufficient thrust or lift had been developed to sustain its flight.

To prevent this from happening, the weapon is restrained on the launcher until sufficient thrust is generated. The restraining device may be simply a pin that is sheared when the weapon develops the required thrust, or it may be a more complicated, reusable device that releases the weapon when the required thrust is exerted.

The launcher must be protected from blast and subsequent corrosive effects of the propellant exhaust stream. This is usually achieved by a protective covering of the launcher and by blast deflection plates or shields. The launcher structure must be able to resist the deteriorating effects of the propellant gases, and the structural design of the launcher should be designed so as to minimize loading or stresses due to propellant blast or recoil momentum.

As with any launcher, the reaction launcher must be compatible not only with space and weight limitations imposed by the delivery vehicle, but also with the environment in which the vehicle operates. Launcher design is further influenced by the degree of control and guidance required of the weapon system. An uncontrolled rocket may require a considerable amount of flight control by the launcher before it is released into free flight, whereas a guided missile does not require this initial control to the same degree. Consequently, reaction launchers are divided into three classes: rail, zero-length, and canister.

Rail Launchers

The term *rail launcher* may be applied to launchers making use of rails, tubes, long ramps, and even tall vertical towers. All provide, to a varying degree, a constraint to the weapon while it is moving along the launcher rail, and thus they provide a considerable amount of flight control. For uncontrolled weapons such as rockets, rails must be relatively long so that the rocket is constrained for a longer portion of the rocket motor burning time to provide the necessary initial velocity vector control.

If the missile is equipped with a guidance system, the rail length can be reduced accordingly. Long-range weapons, guided or unguided, normally require a longer rail since their initial acceleration is comparatively low relative to that of short-range weapons. A rocket booster may also be employed to provide sufficient acceleration and permit reduction of rail length.

Rail launchers may be fixed or movable and are deployed aboard ships, submarines, aircraft, and in the field. They also may serve as ready service storage for missiles and provide facilities for fueling and servicing. The simplicity of the design of rail launchers also promotes reliability and ease of maintenance and repair.

Airborne launcher rails, because of space and airflow problems, are usually very short. As a rule, the greater the speed the missile can achieve at separation from the launcher, the less the airflow effects will influence missile flight.

Thrust developed parallel to the axis of the rail or tube will propel the weapon along the rail and into a proper trajectory. Vertical components of thrust will force the weapon against the rails or tube and thus against the launcher. The launcher must, therefore, be capable of withstanding a portion of the weapon thrust as well as supporting the weapon.

Lateral constraint of the weapon is necessary in order to prevent the weapon from lifting from the rails or wandering during its motion along the launcher rail. This is usually achieved by lugs on the weapon, which ride in slots in the launcher rails. In general, the longer the launcher rails, the better the initial flight control and the less the launch dispersion.

However, long rails, like long beams, can sag and bend. Therefore the longer the beam, the greater the possibility of deflection unless the rail is well supported along its entire length. Because this kind of support is not practical in tactical launchers, a certain amount of beam deflection or droop occurs at the end of long rail launchers.

This droop and the effects of vibrations that are set up as the missile moves along the rail cause the rail to whip and produce unwanted deviations in missile orientation and flight. Thus, material and structural characteristics will limit the length of a rail launcher for tactical usage. This necessary restriction in rail length is not as serious as it might seem, however, because the effectiveness of a rail launcher in constraining weapon travel to the direction of the rails varies with the ratio of rail travel time to the total time of boosted uncontrolled flight.

Because a weapon starts from rest on the launcher, the time of rail travel will vary as the square root of its length. Since

$$S = \frac{1}{2} a t^2 \qquad (14\text{-}4)$$

then

$$t = \frac{\sqrt{2S}}{a} \qquad (14\text{-}5)$$

where:

S = rail length
a = missile acceleration
t = elapsed time of travel

Therefore, a relatively short rail will still provide considerable initial flight control to a weapon.

Zero-Length Launchers

Through common usage, the classification "zero-length launcher" has come to mean any rail launcher where the weapon travels fewer than 8 centimeters before it is released. Since the weapon separates from the launcher shortly, or even immediately, after it begins to move, little or no effective flight control is provided by the launcher.

Therefore, this type of launcher is used with weapons that can immediately assume stable flight. The chief advantages of zero-length launchers are their small size, light weight, comparative simplicity, and ease of maintenance and repair. Also, because of their small size and weight, zero-length launchers require a minimum of deck space and are easily moved in train and elevation. Therefore, they are used extensively in shipboard, ground, and airborne installations for the launching of guided missiles; see figure 14-11.

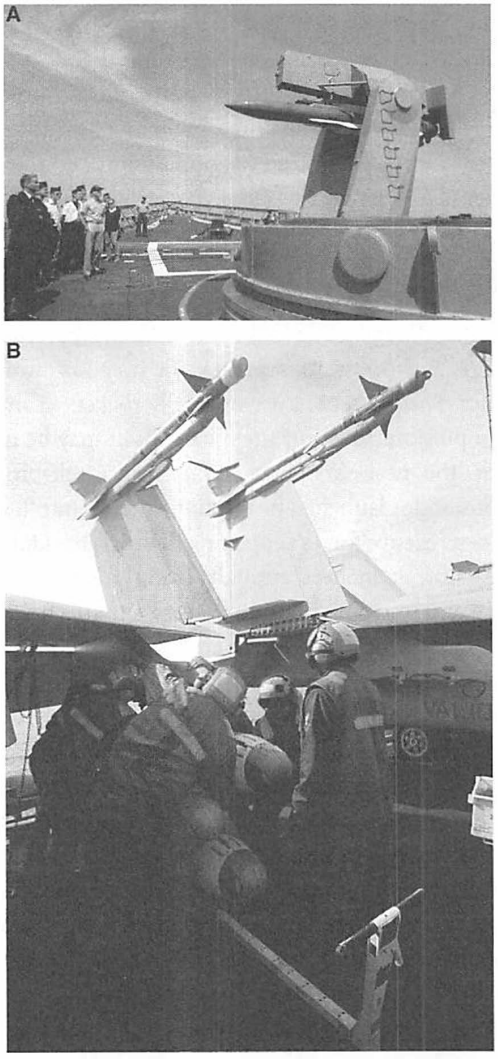


FIGURE 14-11 *Shipboard and aircraft zero-length launchers*

Canister Launchers

With the advent of solid-state guidance, all-electric controls, and solid-fuel rocket motors that have a long

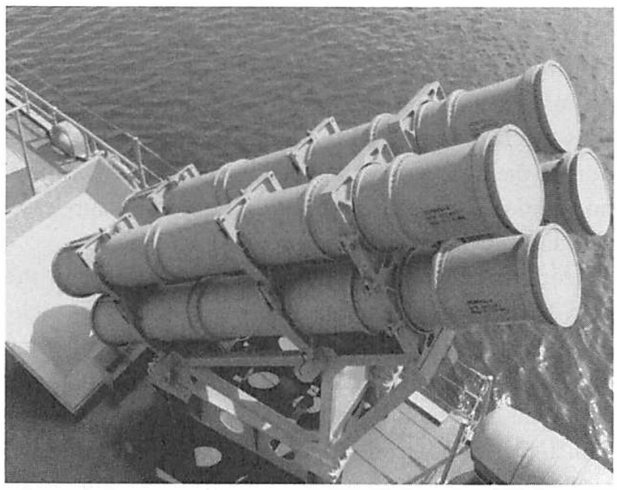


FIGURE 14-12 *Harpoon canisters on a DDG*

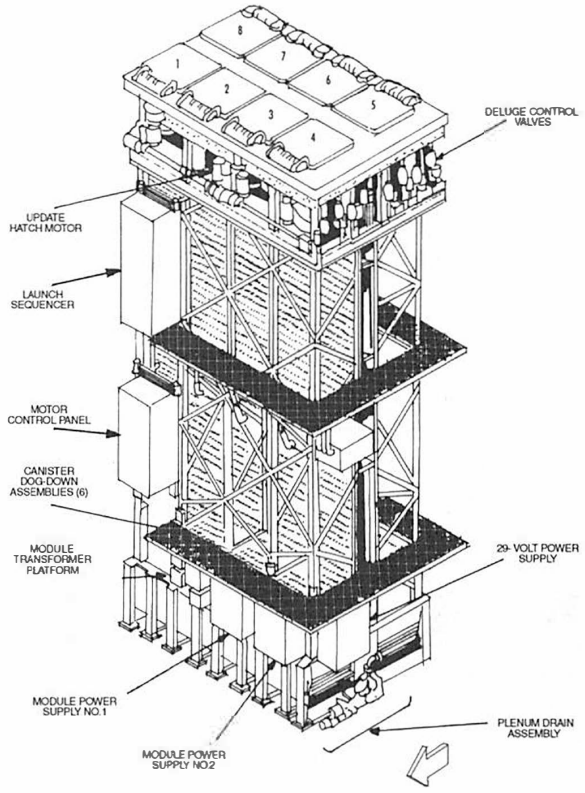


FIGURE 14-13 *Vertical launch system*

service life, guided missile storage requirements have been greatly simplified. These advances, combined with missile flight characteristics that allow the use of zero-length launchers, make it possible to include the shipping container of the weapon as an integral part of the launcher. The resulting canister launcher contains the missile, which is supported by zero-length rail segments or by the surface of the container for lightweight weapons.

The canister may be handheld, as in the Stinger and Javelin missiles. For larger missiles such as the Harpoon, the canister containing the weapon is installed on or within a support structure or launch tube that may be either fixed or mobile; see figure 14-12. In either case, the weapon remains within a sealed container until launch. This minimizes direct handling of the weapon and limits its exposure to adverse environments.

The vertical launch system (VLS), shown in figure 14-13 employs a standard canister that may contain a variety of missile types that are deployed in eight-cell modules aboard ship. The result is a net increase in weapon stowage as well as rate of fire.

Each cell and canister acts as a separate launcher and is independent of the others except for the exhaust gas management system that is common to each eight-cell module. Because there is no launcher reload time, a much higher rate of fire can be maintained.

A disadvantage of the canister launcher is the additional stresses placed on the weapon because of the interior ballistics of the system. The canister confines the weapon for a short period of time after rocket motor light off, subjecting it to shock waves from the flow of high-velocity combustion products prior to the rupture of the canister end closures and the weapon's emergence.

Ignition of the rocket motor causes shock waves to develop between the weapon and the inner wall of the canister. This shock wave is retransmitted from the forward closure back to the missile radome. The net result is a pulsating axial load on the weapon due to dynamic gas phenomena in the canister during launch. The effects of these forces vary with the specific system, but have been known to reach 1,000 newtons in the VLS employing the Standard missile.

In the special case where the canister is stored and fired from within a ship hull, such as in VLS, exhaust gas management is a significant problem. The exhaust stream typically has a temperature of 2,400° Kelvin and a velocity of over 2,500 meters/sec. Additionally, it includes highly abrasive particles and active chemical agents.

Whatever system is used to dissipate the missile exhaust, it must be able to withstand strong pressure waves, not only for the brief time between motor ignition

and weapon departure, but also during the complete burn of the rocket motor with the weapon restrained in the canister in the event of an accident.

The ability to handle sustained rocket motor burn has been a missile magazine design constraint since the time missiles were first placed aboard ship; the intent in VLS, however, is to make sure not only that the ship will survive, but also that the remaining missiles will still be operable.

Steel surfaces are protected from melting or weakening by a replaceable ablative coating that dissipates heat by vaporization during exposure to rocket motor exhaust. The exhaust stream is routed to the atmosphere by a duct system that avoids impingement on unprotected surfaces; see figure 14-14.

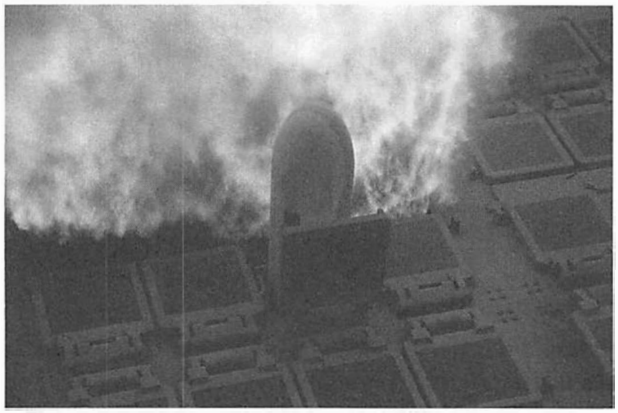


FIGURE 14-14 *Missile launch from a vertical launching system cell*

Launcher Cutouts

Due to obstructions, guns and missile launchers aboard ship and ashore are subject to constraints in their coverage. Shore-based systems must deal with terrain features and man-made structures if a clear, flat emplacement area cannot be found. Shipboard weapon systems must deal with deckhouses, antennas, lifelines, boats, and other weapons in their firing arcs.

All systems employ some type of electromechanical linkage that inhibits the firing circuit at values of bearing and elevation that would cause the weapon to damage structures or be damaged or deflected during launch. These firing cutout mechanisms can also be constructed so that the launcher or gun will automatically elevate to a safe firing elevation at bearings where obstructions are located, while notifying the controlling weapon system of the circumstances.

Gravity Launching Systems

Gravity launchers are characteristically simple in design because they rely upon gravity to separate the weapon from the launcher. Since the initial velocity and orientation of the weapon are supplied by the delivery vehicle, no additional forces are applied to the weapon at release, and the launcher incurs no sudden shock.

Airborne platforms used this type of system in the past. However, with the great speeds obtained by high-performance aircraft, the use of gravity launching for any but small practice weapons has become unfeasible.

Because of the airflow patterns about the weapon and launch platform, aerodynamic forces are created that can result in the weapon not separating from the aircraft or separating and then striking the aircraft. Therefore, pure gravity launching systems are seldom used.

Functions of a Launching System

Any launching system must perform certain distinct functions in order to successfully launch the weapon associated with it. In the operational configuration, these functions do not necessarily occur in isolation, but instead frequently integrate as processes in achieving the intended goal of the system. These basic functions are as follows:

1. *Storage.* A safe and readily accessible area must be provided for the storage of weapons until needed.
2. *Transfer.* The weapons must be moved from the storage position to the loading position and vice versa.
3. *Loading.* Before the weapon may be put into flight, it must be placed on the launching device in a ready-to-fire position.
4. *Control.* Once the weapon has been loaded, it must be oriented in space. Typically a launching system functions in response to the solution of the fire control problem.
5. *Launching.* Flight initiation occurs during the process of launching as a result of a propelling force provided to the weapon by gravity, an impulse, or a reaction.

Storage

To achieve and sustain optimum firing, the launching system selected for a given weapon system must include a storage subsystem capable of performing its allotted functions in a minimal period of time. The capacity of the subsystem must be adequate to sustain the desired firing operation, and the facilities must be readily accessible. A *magazine* is a term used to denote a storage area needed

to provide the space and the safety facilities for storage of large quantities of gravity, gun, or reaction-type weapons. Magazines are classified in three categories.

Primary magazines These magazines are designed to stow the delivery vehicle's complete peacetime allowance of ammunition. Primary magazines are typically well protected and equipped with thermal insulation, ventilation, and sprinkling systems designed to cool the weapons below an established upper temperature limit. A method of ensuring both physical and environmental security, through the use of locking devices and remote alarm systems, for example, is associated with the primary magazine.

In the case of reaction-type weapons, restraining equipment for the weapon is incorporated to prevent an ignited propulsion motor from actually moving the weapon. Vents are also provided to prevent disruptive pressures from developing within the magazine in the event of inadvertent propulsion motor ignition.

Ready service lockers This type of storage designates the ammunition storage facilities in the immediate vicinity of the launcher. This storage is used only when ammunition is required to immediately service the launcher. Under this magazine classification scheme, certain gravity and reaction-type launching systems do not have primary magazines as such, but may be said to have only ready-service magazines. However, the characteristics of primary magazines are normally incorporated into the design of these storage facilities.

Lockers These are compartments designed for storing sensitive or special types of ammunition such as fuzes, pyrotechnics, blasting caps, hand grenades, and explosive chemicals. Lockers are normally located above the damage control deck and on a weather deck if possible in order to localize and minimize the effects of fire or other adverse conditions.

Most magazines are single-purpose—that is, they are designed for a homogeneous quantity of ammunition. Multipurpose magazines may be required because of space limitations in the delivery vehicle. In this case, different types of ammunition will be stored. Such mixed storage will not include special types of ammunition designated for locker storage.

Where mixed storage is employed in the multipurpose magazine, the various types of ammunition are segregated for ease of identification and accessibility. Very specific requirements exist specifying ammunition types that may be stored in the same magazine or locker and those that are incompatible.

Transfer Equipment

Weapon transfer equipment is designed to move weapons and associated components—fuzes and special payloads—from the primary magazine to ready-service lockers and then as needed to the launcher. In addition, the transfer system is normally capable of returning the weapon to the magazine.

To achieve rapid initial employment, the transfer system must be capable of moving weapons at a rate commensurate with the launcher rate of fire. If a transfer line or channel has a transfer rate less than the required rate of fire, then two or even three transfer lines may be necessary to feed the launcher.

When transferring weapons to rotating launchers, it is necessary at some point to shift from nonrotating transfer equipment to equipment rotating with the launcher. This becomes a problem of considerable magnitude for a high-rate-of-fire system where the weapon must be transferred while the launcher is rotating. This can be accomplished by a manual transfer or by complex automatic equipment, as in high-rate-of-fire naval guns. When the weapons to be transferred are large, it is often necessary to have the rotating launcher return to a specified position before a weapon is transferred to it.

Weapon flight preparation facilities are often found in the transfer system. En route to the launcher, operations such as weapon assembly, checkout, servicing, and even programming of the weapon may be performed.

To prevent danger to operating personnel and damage to the delivery vehicle in the event of weapon explosion or ignition, the transfer system must be designed to localize and contain detonations, fires, propulsion motor blasts, and so on. In addition to lethal effects of the weapon itself, personnel can be injured by the moving and rotating machinery of the launching system. Thus, in the transfer system as well as in every component of a launching system, careful attention must be given in the design of equipment to minimize the danger of accidents.

The transfer of ammunition can be as simple as moving an airborne missile on a dolly from the primary magazine, via elevators, to the launcher mounted on an aircraft, or it can be as complex as a system that provides for the automatic transfer of a large weapon weighing in excess of 1 ton to a torpedo tube.

In the more complex transfer systems, hydraulic, pneumatic, and electrical servo systems are employed to accomplish the task in such a manner that operating personnel can monitor system functions from only a local or remote control panel.

Loading Equipment

Weapon loading equipment is needed to place the weapon in firing position on the launcher in a rapid, reliable, and safe manner. The loading operation consists of moving weapons from the transfer equipment to the launcher and positioning them on the launcher.

Transfer and ramming functions can be performed manually, mechanically with human operator control, or automatically with human monitoring. In some types of fixed launchers with low rates of fire, the ramming operation is not necessary. The weapon is simply hoisted or lowered into position on the launcher. However, loading high-rate-of-fire launchers requires a fast, precise ramming cycle.

Some means also must be provided for unloading unfired but usable weapons and for jettisoning unsafe and dud weapons. In most cases, loading equipment is designed to accomplish these functions when it is not desirable for personnel to accomplish this. In some cases, special equipment is provided to perform the jettisoning function.

To achieve rapid weapon employment and to maintain a high rate of fire, the loading rate must be compatible with the required launcher rate of fire. As with transfer equipment, this often requires the use of more than one loading unit per launching tube or rail, or more than one line or channel feeding weapons to the ramming equipment.

The majority of automatic loading systems employ either a catapulting device, a rail spade, or a chain device.

Rail-spade ramming This method is frequently employed in the ramming of semifixed (projectile and powder separate) ammunition and in loading of torpedoes and reaction weapons where the weapon is housed in a ready-service magazine.

The size of many reaction weapons requires the use of manually controlled loading equipment. In this case, the weapon is supported by the loader, which may be fixed or portable, in the vicinity of the launcher and is slowly rammed into position. This is a comparatively slow process that is suitable only for launchers that are not expected to engage a large number of targets between reloading.

Catapult ramming This method is found in gun launchers that employ fixed ammunition. A round is received from the transfer system and aligned with the gun bore. The tray in which the round is resting is then rapidly moved toward the breech, and abruptly stopped. At the end of the tray's travel, the round is released and catapulted into the chamber, whereupon the breechblock seals the breech. Figure 14-15 portrays this loading method.

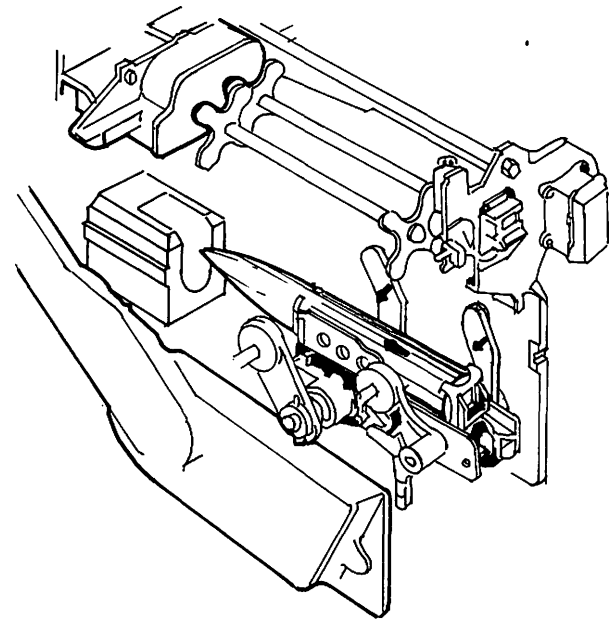


FIGURE 14-15 *Catapult ramming of field ammunition in a gun launcher*

Chain ramming A flexible-link chain is frequently employed in the previous two categories of ramming, but in this category, the end of the chain is provided with a buffer pad that engages the base of the weapon. The chain is driven by a sprocket along a track, where it is free to move forward or retract, as required. In figure 14-16, notice that the chain can flex in only one direction. The buffer serves to protect the weapon from damage since velocities of more than 3 meters/sec can be achieved by this ramming method in some designs.

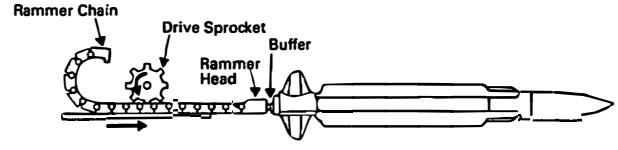


FIGURE 14-16 *Chain ramming of a guided missile in a reaction launcher*

Control

Once the weapon has been loaded in the launcher, the final step of engaging a target takes place—delivery of the weapon. For this to occur, the launcher must be positioned along a line that will ultimately result in the weapon intercepting the target. This is, appropriately, called the line-of-fire (LOF).

In general, the launching system is subordinate in weapon-system dynamics in that it must respond to certain types of orders. Depending upon the weapon system

employed, these orders are known as gun orders, launcher orders, torpedo orders, missile orders, and so on. The specific nature of these orders will vary from system to system.

For instance, a typical set of gun orders may be: gun train (Bdg'), elevation (Edg'), sight angle (Vs), sight deflection (Ls), parallax (pl), and fuze time (T5). The alphanumeric symbols associated with each of the orders are known as fire control symbols. They are often employed as a shorthand device for describing information/data flow in weapon systems.

The following is an example of a set of missile orders: seeker head orders, Doppler predict, ship roll correction, channel select, gravity bias, and target select. Each of these orders will be continuously transmitted from the computational system to the launching system and/or to the weapon itself during the interval between launcher assignment and weapon launch or until the launcher has disengaged electronically.

A weapon system will often have several types of weapons available to launch, so it must receive orders regarding the type or types of weapons to be fired. For example, in a shipboard gun system, several types of projectiles are in the magazines. Control station personnel must select the type of projectile to be fired and notify the launching system of their choice. For a shipboard missile system, the launcher must receive orders regarding which type of missile to launch.

In addition to the type of missile, the launching system must know how many missiles to launch. For example, a submarine launching system may fire one, two, or a whole spread of torpedoes. A guided-missile launching system may fire a one- or two-round salvo, and a gun system may fire continuously until a target is destroyed or out of range.

However, the most important function of the launcher is to provide initial flight orientation to the weapon. In cases where the launcher is fixed rigidly to the delivery vehicle, the whole vehicle must be maneuvered to provide this orientation. If a launching system is mounted in a vehicle that is not very maneuverable, the launcher must be rotated about one or more axes.

As discussed in previous chapters, a sensor system will detect, classify, and locate a target such that present target position is continuously entered into the computational system. It is the output of the computational system that provides the orders that cause the movement of the launcher to the LOF. As discussed in chapter 13, once the line-of-sight (LOS) is defined by the tracking system, the weapon launch lead angle is computed. The gross measure of this angle is between the LOS and the LOF.

Movement of the launcher in bearing (train) and elevation to the LOF is accomplished by respective analog signals in the form of DC error voltages or digital signals representing magnitude and direction. The launcher is normally referenced to the delivery vehicle as discussed in weapon systems alignments in chapter 13. The error signals are entered into the train and elevation servo systems of the launcher. The resulting drive signals cause the launcher to respond in train and elevation, while employing position and velocity feedback in a manner identical to that described in chapter 12.

In order to control a launching system effectively, a scheme of communications must be incorporated. Physical movement of the launcher and transfer/loading functions may be directed by simple voice commands, followed by an appropriate action response.

In more complex systems where multiple weapons are employed against various, simultaneous targets, communications are accomplished by information systems using electronic data-processing components. For extremely complex weapon systems such as Aegis, communication and control may be initiated by the operator, after which the complete cycle of transfer, loading, positioning, and launching will occur without further human operator input.

Launching Systems Summary

In the case of impulse and reaction systems, the launching of the weapon is usually brought about by the activation of the propelling charge in such a manner that expanding gases result in the expulsion of the weapon. Launching for gravity systems is brought about by the simple release of the weapon.

Launching systems must provide for the safety of operating personnel and the vehicle in which the systems are installed. A great number of safety features are included in the various components that make up the entire system. Most launchers are equipped with mechanisms that prevent the launcher from pointing or firing at the delivery vehicle.

Mechanical devices—such as firing cutout cams and electrical circuitry programmed to cause arbitrary launcher movement upon approaching blind or nonfiring zones—are employed to ensure that the line-of-fire never occurs within a danger zone without warning to operating personnel.

Methods to evacuate or redirect propellant gases are a design feature of launchers to prevent physical harm to personnel and damage to components (gas ejectors in guns, blast deflectors and diffusers in reaction launchers).

Generally, the delivery vehicle is protected from exhaust damage by the design and location of the launcher on the vehicle as well as by securing an unobstructed area around the launcher and using protective materials.

Weapon Propulsion

The power required to propel a warhead to its target is obtained through the controlled release of stored energy. Usually types of propulsion are considered from the viewpoint of the energy source. The energy source may be the end product of:

1. A chemical reaction
2. A compression of gases or liquids
3. The effect of gravity

Impulse Propulsion

Impulse propulsion systems include all weapon systems in which a projectile is ejected from a container, usually a long tube, by means of an initial impulse. Also included in this classification are systems that employ compressed gases to provide the initial impulse propulsion until the systems have cleared the launching vehicle—for example, torpedo and submarine-launched ballistic missile systems. Potentially in the future, electromagnetic force might provide the impulse.

The expulsion of projectiles at the high velocities demanded by modern warfare requires tremendous forces. The study of these forces and of the action within a gun at the time of fire is referred to as interior ballistics, discussed in chapter 13. It comprises a study of a chemical energy source, a working substance (high-pressure gas), and the equipment to release and direct the working substance. To provide the high response speed needed in a gun, the propellant must transfer its energy to the projectile by means of the expanding gaseous products of the combustion.

The gas that causes the pressure that propels the projectile is generated by the ignition of an explosive train. This explosive train is termed a propellant train and is similar to the high-explosive train that is discussed in chapter 16. The difference is that the propellant train consists primarily of low level explosives instead of high level explosives and has a primer, an igniter or igniting charge, and a propelling charge.

The ignition of a small quantity of sensitive explosive—the primer (for example lead azide)—is initiated by a blow from the firing pin or electrical shock and is transmitted and intensified by the igniter so that the large, relatively insensitive propelling charge burns in the proper manner and launches the projectile; see figure 14-17.

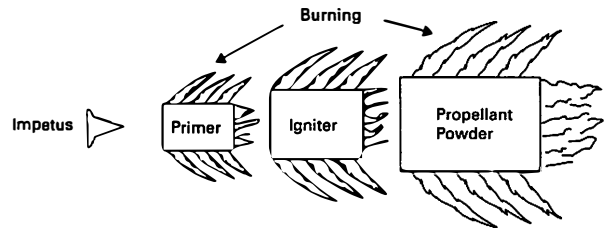


FIGURE 14-17 *Explosive propellant train*

Reaction Propulsion

Contrary to popular belief, reaction motors do not obtain thrust by "pushing" against the medium in which they are operating. Thrust is developed by increasing the momentum of the working fluid and creating a pressure differential therein. Newton's Third Law of Motion, which states that for every action, there is an equal and opposite reaction, forms the basis for motion of all self-propelled objects.

Pressure differentials are employed to move propeller-driven aircraft, rockets, and jet-propelled missiles. By definition, a reaction-propelled vehicle is one that contains within its structure its own source of propulsion—that is, a reaction-type motor. The mediums in which reaction motors operate and the types that have been used in weapon systems are:

1. Air: propeller engines, turbojets, ramjets, rockets
2. Vacuum: rockets
3. Water: screws, hydrojets, rockets

The performance required of modern propulsion systems, insofar as range, velocity, and control are concerned, is considerably greater than that which can be accomplished by once-conventional methods. Until the late 1950s, the reciprocating engine-propeller combination was considered satisfactory for the propulsion of aircraft. But the shock waves generated as the speed of sound is approached limit the development of thrust and thus the speed of propeller-driven craft.

Although future developments may overcome the limitations of propeller-driven vehicles, it is necessary at the present time to use jet propulsion for missiles traveling at high subsonic and supersonic speeds. A minor disadvantage of jet propulsion is that tremendous quantities of fuel are needed per second of flight time. Also, the only means of self-propulsion that will operate in a vacuum is propulsion with rocket motors.

Classification of Jet-Propulsion Engines

Jet propulsion is a means of locomotion obtained from the momentum of matter ejected from within the propelled body in the form of a fluid jet. The fluids used for producing the jet include water, steam, heated air, and gases produced by a chemical reaction. The fluid may be drawn into the body from the very medium in which the body is moving, such as air in a jet engine or water in a hydro-jet engine; or it may be self-contained, as in the case of rocket engines, where the working fluid is composed of gaseous products of rocket fuel combustion.

A jet-propulsion engine consists essentially of an air inlet section (diffuser) for air-breathing jets, a propellant supply system, a combustion chamber, and an exhaust nozzle. The purpose of the propellant system and combustion chamber is to produce large volumes of high-temperature, high-pressure gases. The exhaust nozzle then converts this heat energy into kinetic energy as efficiently as possible.

Conventional terminology makes a distinction between jets and rockets: A jet takes in air from the atmosphere; a rocket needs no air supply because it carries its own supply of oxygen. Both types of engines operate by expelling a stream of gas at high speed from a nozzle at the after end of the vehicle. For our purposes here, a rocket can be considered a type of jet engine.

In a thermal jet engine (air-breathing), the missile brings in a quantity of air at its input end and compresses it. Liquid fuel is then injected into the compressed air and the mixture is ignited in a combustion chamber. As the resulting hot gases are expelled through a nozzle at the rear of the vehicle, heat energy is transformed into kinetic energy, and thrust is created.

As previously mentioned, the basic difference between a rocket and a thermal jet engine is that a rocket carries its own oxidizer and needs no outside air for the combustion process. In addition, thrust developed by an air-breathing engine is dependent on the momentum change between the fluid (air) entering the engine and the fluid (hot gases) exiting the engine, whereas in a rocket, thrust depends only on the momentum or velocity of the jet of exhaust gases.

Furthermore, the thrust developed by a rocket per unit-of-engine frontal area and per unit-of-engine weight is the greatest of any known type of engine. In liquid-fuel rockets, fuel is pumped into the combustion chamber to be burned, whereas in a solid-propellant rocket, the combustion chamber already contains the fuel to be burned.

Rockets are distinguished by the means they use to produce exhaust material. The most common type of rocket engine obtains its high-pressure gases by burning a propellant.

Propellants

The fuels and oxidizers used to power a jet/rocket engine are called propellants. The chemical reaction between fuel and oxidizer in the combustion chamber of the jet engine produces high-pressure, high-temperature gases. These gases, when channeled through an exhaust nozzle, are converted into kinetic energy, which create a force acting in a direction opposite to the flow of the exhaust gases from the nozzle. This propulsive force, termed thrust, is a function primarily of the velocity at which the gases leave the exhaust nozzle and the mass flow rate of the gases.

In order to develop a sufficient thrust with a solid propellant, grains or charges of propellant with large burning surfaces are employed so that a high rate of mass flow is developed. The duration of the burning of a propellant charge is determined by the configuration of the grain and its burning rate. Since the combustion chamber has fixed dimensions and a fixed capacity for propellant thrust, it may be either great but of short duration, or low and of long duration.

The thrust developed by a reaction motor is the result of static pressure forces acting on the interior and exterior surfaces of the combustion chamber. This force imbalance is shown graphically in figure 14-18.

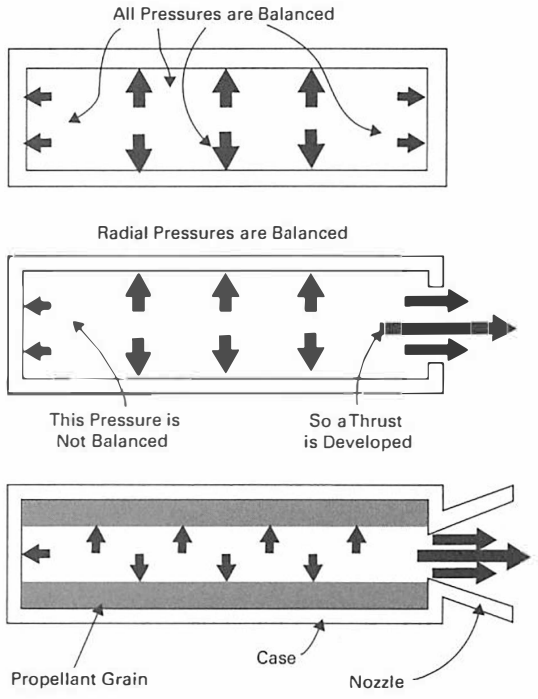


FIGURE 14-18 *Development of thrust in a rocket motor*

Static pressures acting on the interior surfaces depend in turn on the rate at which propellants are burned, the thermochemical characteristics of the gases produced by combustion, and the area of the throat of the dispersing nozzle. Because the internal forces are several times greater than the external forces, and because the forces acting normal to the longitudinal axis of the combustion chamber do not contribute to the thrust, the thrust developed is primarily the resultant axial component of the pressure forces.

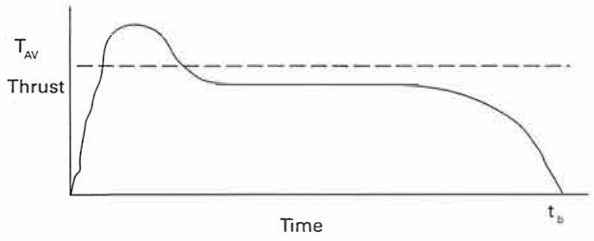


FIGURE 14-19 *Thrust-time curve*

If the thrust-time curve (see figure 14-19) obtained from firing a rocket is integrated over the burning duration, the result is called total impulse and is measured in newton-seconds. Then

$$I_t = \int_0^{t_b} T\,\partial t = T_{av} t_b \qquad (14\text{-}6)$$

where:

I_t = total impulse in newton-seconds (N-sec)
t_b = burn time (sec)
T_{av} = average thrust in newtons (N)

The performance characteristics of a specific propellant are called the specific impulse of the propellant (solid-fuel) or specific thrust of the propellant (liquid-fuel). The specific impulse is a measure of the "quality" or "merit" of the fuel and is defined as:

$$I_{sp} = \frac{T_{av} t_b}{m_p g} = \frac{I_t}{m_p g} \qquad (14\text{-}7)$$

where:

I_{sp} = specific impulse (sec)
m_p = mass of solid propellant (kg)
g = force of gravity (9.8 m/sec^2)

Although the units of specific impulse are seconds, the I_{sp} of a fuel is actually the amount of impulse per kilogram of fuel. In other words, the specific impulse reflects the specific energy of the fuel. Thus, the fuel with the highest I_{sp} will produce the greatest performance.

Propellants are classified as either solid propellants or liquid propellants. Nearly all rocket-powered weapons in use by the U.S. Navy use solid propellants. Liquid propellants are still used in some older ICBMs and are often used in cruise missiles. Of course, all thermal jet engines burn a liquid fuel.

Solid propellants consist of a fuel—usually a hydrocarbon—and an oxidizer, which are combined to produce a solid of desired chemical potential. Burning rate, which directly affects the amount of thrust produced, is the velocity in meters per second at which the grain is consumed. Burning rates of propellants depend upon the chemical composition of the propellant, the combustion chamber temperature and pressure gradient, the gas velocity adjacent to the burning surface, and the size and shape of the grain or the geometry of the overall propellant.

Rocket propellants are cast to obtain the types of burning previously discussed—in chapter 13 namely, neutral, progressive, or degressive burning. Propulsive specifications will determine which type is to be used. Normally a uniform burning rate is desired so that a constant thrust is produced. Most U.S. Navy propellants use either a cruciform grain or a cylindrical grain with an axial hole and radial perforations; see figure 14-20.

The cruciform grain in cross section is symmetrical. If all of the exterior surface of this grain were permitted to burn, there would be a gradual decrease of area, and the burning rate would be degressive. Since a uniform burning

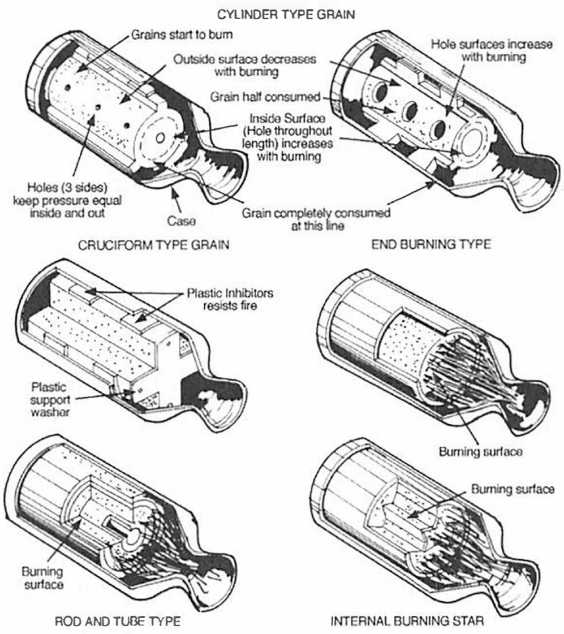


FIGURE 14-20 *Solid-propellant configurations*

rate is desired, a number of slower-burning plastic strips or inhibitors are bonded to certain parts of the area that are exposed on the outer curved ends of the arms. These control or slow the initial burning rate, causing the gas-production rate to be approximately uniform over the burning time. Other cylinder grain types are also shown in figure 14-20.

Liquid propellants are normally stored in tanks outside the combustion chamber and are injected into the combustion chamber. The injector vaporizes and mixes the fuel and the oxidizers in the proper proportions for efficient burning. Liquid propellants are classified as either monopropellant or bipropellant. In the monopropellant type the fuel and the oxidizer are mixed, and in the bipropellant type, the fuel and the oxidizer are separated until firing.

Liquid fuels are more powerful than solid fuels; but other than this advantage, a liquid-fuel rocket is not ideally suited as a weapon-propulsion system. Because of their high volatility and corrosive nature, liquid fuels cannot be stored for long periods of time, which usually means that the system must be fueled just prior to launch. This negates its ability to be a quick-reaction weapon, which is usually required in combat situations.

Solid-fuel rocket engines do not require complex plumbing systems and are fairly simple systems, as illustrated in figure 14-21. Storage problems are minimized, and the propellant is usually very stable. For these reasons, solid-fuel rockets are almost exclusively used.

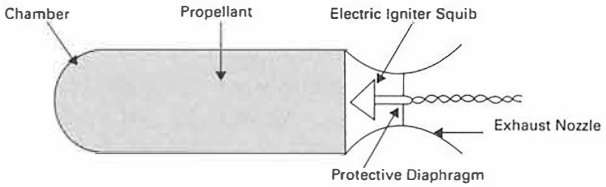


FIGURE 14-21 *Elements of a solid-fuel rocket motor*

Elements of Propulsion Subsystems

A jet/rocket-propelled engine is basically a device for converting a portion of the thermochemical energy developed in its combustion chamber into kinetic energy associated with a high-speed gaseous exhaust jet. The basic elements of the engine are:

1. The combustion chamber, in which the transformation of energy from potential to heat occurs
2. The exhaust nozzle, in which thermochemical energy is converted into the kinetic energy necessary to produce an exhaust jet of propulsive potential
3. The diffuser (for air-breathing jets only) or intake duct, in which the high-speed air intake is converted into low-speed, high-pressure air for entry into the combustion chamber as an oxidizing agent

The combustion chamber is the enclosure in which high-temperature, high-pressure gases are produced and potential energy is converted to kinetic energy. In liquid-fuel engines, injectors bring the fuel and oxidizers together in the chamber. In solid-fuel engines, the propellant is already contained in the chamber. Also contained in the combustion chamber is some type of ignition device to initiate the burning of the propellant.

An exhaust nozzle is a mechanically designed orifice through which gases generated in the combustion chamber flow to the outside. The function of the nozzle is to increase the exit velocity of hot gases flowing out of the engine so that maximum thrust is extracted from the fuel. A nozzle consists of a mouth, throat, and exit, as illustrated in figure 14-22.

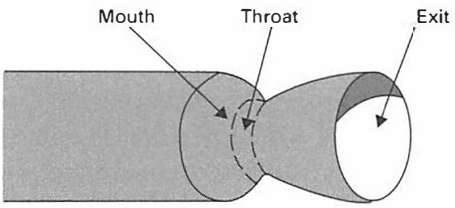


FIGURE 14-22 *Nozzle components*

Under conditions of steady flow, the mass flow rate, according to Bernoulli's theorem, remains constant. Thus, in subsonic flow, where the density of the gas/fluid is considered to remain constant, velocity must increase at any point where the cross-sectional area decreases in order to accommodate a constant mass flow rate. If flow remains subsonic, then as the cross-sectional area increases, the velocity will decrease.

It is possible to reduce the cross-sectional area of the nozzle throat to the point that the velocity will reach a maximum and become sonic. Then, with proper design of the divergent section of the nozzle, the flow velocity can be expanded supersonically as the cross-sectional area increases.

Since thrust from a rocket motor is proportional to the momentum of the exhaust gases ejected per second, and momentum is equal to mass times velocity, thrust efficiency can be increased at no extra cost in fuel consumption if the exhaust velocity is maximized by proper nozzle design.

One of the more efficient designs in common use is the De Laval nozzle. The De Laval nozzle converges to a throat at which the velocity becomes sonic, and then the nozzle diverges such that the flow is expanded supersonically.

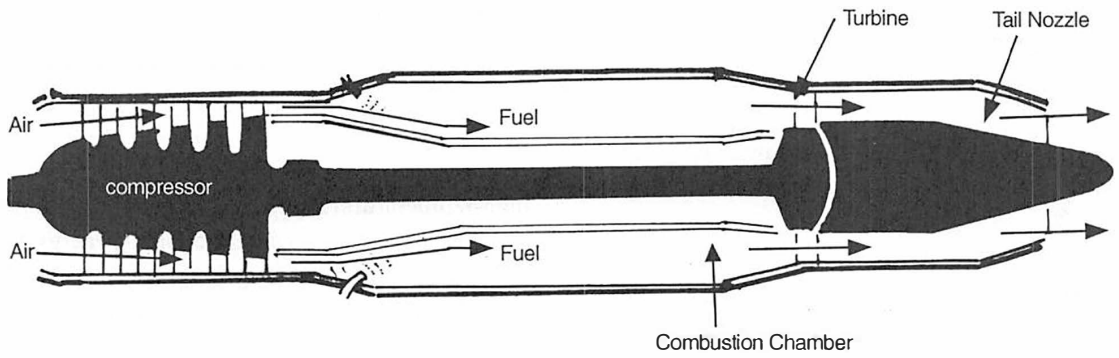


FIGURE 14-23 *Axial-flow turbojet*

Turbojet Engines

A turbojet engine derives its name from the fact that a portion of its exhaust output is diverted to operate a turbine, which in turn drives the air compressor that is used to compress the input airflow. Modern turbojets are of the axial-flow design as depicted in figure 14-23. An axial-flow compressor is similar in operation to a propeller except that is has multiple fixed (stator) and rotating (rotor) blades. As the rotor of the axial compressor turns, the blades impart energy of motion in both a tangential and axial direction to air entering the front of the engine from the diffuser.

The function of the diffuser or intake duct is to decelerate the velocity of the air from its free-stream velocity to a desired velocity at the entrance of the compressor section or combustion chamber, with a minimum of pressure loss. For both subsonic and supersonic diffusers, flow is decelerated from the inlet to the engine compressor section.

It is important to note that airflow must be subsonic before it enters the engine. The stator is set in a fixed position with its blades preset at an angle such that the air thrown off the first-stage rotor blades is redirected into the path of the second-stage rotor blades. The added velocity compresses the air, increasing its density and, as a result, its pressure. This cycle of events is repeated in each stage of the compressor. Therefore, by increasing the number of stages, the final pressure can be increased to almost any desired value.

The combination of the air-intake system, air compressor, combustion system, and turbine is essentially an open-cycle gas turbine combined with a jet stream. In operation, the compressor, driven by the gas turbine, supplies air under high pressure to the combustion chamber, where the fuel and air are mixed and burned. The turbine absorbs only part of the energy, and the remainder is employed for the production of thrust.

Once started, combustion is continuous. Various designs include one-, two-, and even three-stage compressors, accompanied by one-, two-, and three-stage turbines to power the compressors. The most common design employs a two-stage compressor.

Inlet and nozzle design is critical to the performance of all turbine engines. It is even more critical to aircraft expected to operate at speeds beyond Mach 1. Without proper intake or diffuser and nozzle design, a jet-propelled aircraft is limited to less than the speed of sound. An improperly designed intake would cause shock waves to develop on the compressor blades and severely interfere with its operation.

Some aircraft are able to utilize a simple fixed inlet that enables speeds in excess of Mach 1, but less than Mach 2. Speeds beyond this usually require variable-geometry inlets where the shape of the inlet can be changed with changes in velocity.

Also essential to supersonic flight is a variable-geometry exhaust nozzle. Just as the shape of the intake slows air and increases pressure, exhaust nozzles trade pressure for speed and accelerate the exhaust gases.

There are several variations of the basic turbojet engine design: turboprop, turboshaft, and turbofan. Turboprop and turboshaft engines have a shaft attached to the second turbine stage, called the power turbine, which turns a propeller (turboprop) or helicopter rotors (turboshaft). The direct thrust produced by the flow of exhaust gases rearward is converted to a torque that drives the propeller or rotors.

A turbofan engine bypasses a portion of the air from the first compressor or fan stage and directs it aft, bypassing the combustion chamber and turbine. Sometimes the bypassed air is mixed with the air from the turbine. There are low- and high- bypass-ratio turbofans. Low-bypass turbofans are typically found on high-performance military aircraft such as fighters. High-bypass turbofans are usually found on large transport aircraft such as airliners. Turbofans are generally more thermodynamically efficient and versatile than turbojets and have for the most part replaced the turbojet in military and commercial service.

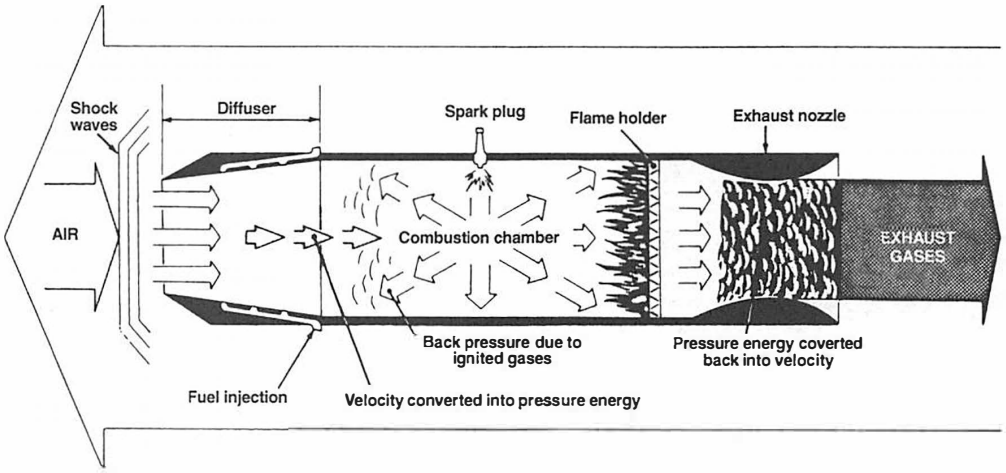


FIGURE 14-24 *Low-supersonic ramjet*

Currently, several missiles use turbojets and turbofans for their main propulsion plants. These are generally longer-range missiles such as the Harpoon, Tomahawk, and the air launched cruise missile (ALCM).

Ramjet Engines

The most promising jet engine, from the standpoint of simplicity and efficiency at supersonic speed, is the ramjet, so called because of the ram action that makes its operation possible.

The ramjet, illustrated in figures 14-24 and 14-25, has no moving parts. It consists of a cylindrical tube that is open at both ends, with an interior fuel-injection system. As the ramjet moves through the atmosphere, air is taken in through the front of the diffuser section, which is designed to convert high-speed, low-pressure airflow into low-speed, high-pressure airflow. Thus, a pressure barrier is formed and the escape of combustion gases out of the front of the engine is prevented.

The high-pressure air then mixes with fuel, which is sprayed continuously into the engine by fuel injectors. Burning is initiated by spark action, after which it is uninterrupted and self-supporting—requiring no further spark assistance. The flame front is kept from extending too far toward the rear of the engine by a device called the flame holder. By restricting the burning to the combustion chamber, the flame holder maintains the combustion chamber temperature at a point high enough to support combustion.

The combustion gases bombard the sides of the diffuser and the ram air barrier, exerting a force in the forward direction. Since the gases are allowed to escape only out of the rear through the exhaust nozzle, the force in the forward direction is unbalanced. The degree to which this force is unbalanced depends on the efficiency with which the exhaust nozzle can dispose of the rearward-moving combustion gases by converting high pressure into velocity.

The main disadvantage of ramjet engines lies in the fact that they are, by the nature of their operation, unable to develop thrust at rest (static thrust). If fired at rest, high-pressure combustion gases would escape from the front as well as from the rear. Consequently, before a missile with a ramjet engine can function properly, it must

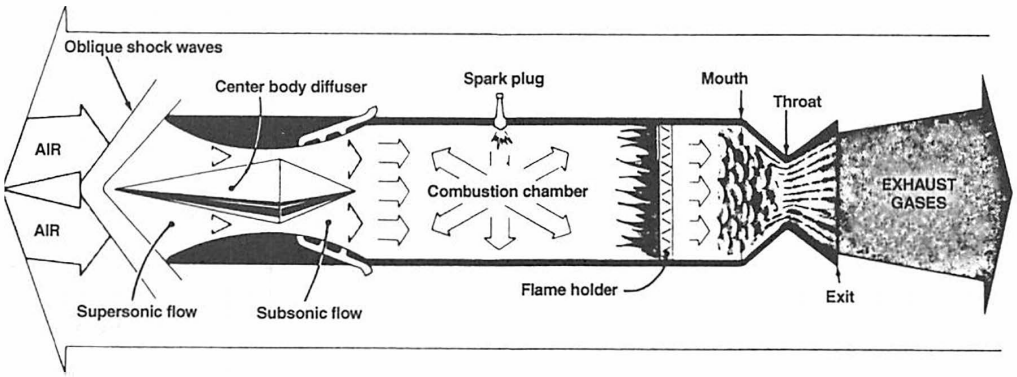


FIGURE 14-25 *Hypersonic ramjet*

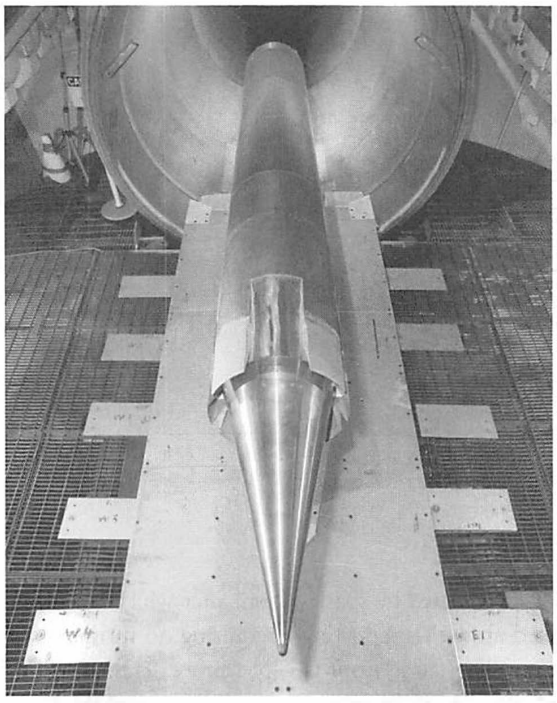


FIGURE 14-26 *A hypersonic cruise missile engine used in the first-ever ground test of a full-scale, fully integrated hypersonic cruise missile using conventional liquid hydrocarbon fuel*[i]

first be boosted by some other propulsion system to a speed approximately equal to that at which the ramjet was designed to operate. Ramjet engines are also restricted to use in altitudes below approximately 90,000 feet, since they must have air to operate.

A new variant of the ramjet, an integral rocket/ramjet, is under development for advanced cruise missiles. The rocket propellant is carried in the combustion chamber of the ramjet and provides the boost to get the ramjet up to high speed. When the propellant has burned, leaving the chamber empty, the ramjet begins operation.

By burning a solid fuel that contains an insufficient oxidizer and combining the oxygen-deficient combustion products with air from outside the missile, engineers have been able to build a solid fuel ramjet. This is especially promising when the advantages of solid fuel are combined with the increased range potential gained from using ambient air rather than an onboard oxidizer. The solid-fuel ramjet shows great potential for increasing both the speed and the range of surface-to-air and cruise missiles.

Ramjets can achieve very high speeds, up to Mach 4, but a new, enhanced kind of ramjet, the scramjet, can go even faster. Scramjet is a semiacronym for supersonic combustion *ramjet*. It works basically like a ramjet, but is designed for supersonic flow in the combustion chamber, allowing the scramjet to fly faster than Mach 5. See figure 14-26.

Gravity-Type Propulsion

Gravity-type propulsion is the propulsion system used for free fall and glide bombs, and for missiles after their reaction propulsion system has ceased functioning. In either case, the missile or bomb has some initial velocity given to it from its launching platform—for instance, for a bomb it would be the aircraft velocity. Once it has been released, the only forces acting on it would be gravity, drag, and lift in the case of a glide bomb. The potential energy of the weapon (its altitude) is converted to kinetic energy (velocity) as it falls due to the force of gravity.

Weapon Architecture

The structure of a weapon is designed to support and protect the warhead, guidance, and propulsion subsystems. It must withstand environmental forces as well as accelerations imposed by weapon motion. It must be lightweight to attain useful velocity easily and have adequate range, while having sufficient strength to resist bending or outright structural failure during high-speed turns.

Additionally, its physical shape must be aerodynamically or hydrodynamically sound to ensure efficiency, while being compatible with storage facilities, launching mechanisms, and platforms. The architecture of the weapon can, to a large measure, determine its success or failure.

Architecture of Missiles

The size and type of a missile selected for a particular function are based on the target, the launch vehicle or platform, range and maneuverability requirements, altitude

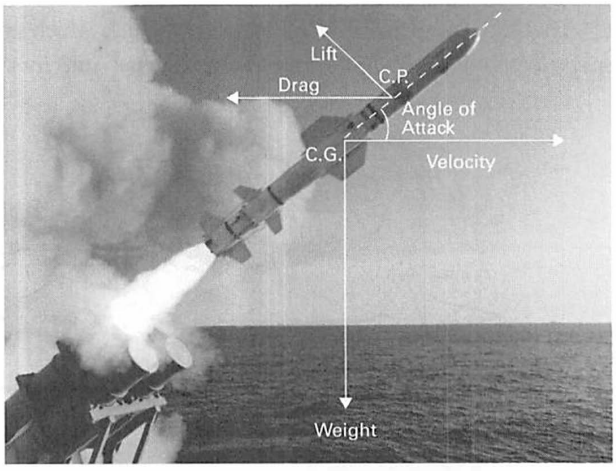


FIGURE 14-27 *Missile force diagram*

TABLE 14-1 *Reaction propulsion advantages by type*

Solid-Fuel Rocket	Liquid-Fuel Rocket	Ramjet	Turbojet
Very simple	Relatively simple	Very simple	Develops large static thrust
Unlimited speed	Practically unlimited speed	No wearing parts. Gets oxygen from air	Consumes fuel only. Gets oxygen from air
Operates in any medium or in a vacuum	Operates in any medium or in a vacuum	Relatively inexpensive to construct and operate	Thrust practically independent of speed
Full thrust on takeoff	Relatively few moving parts	Lightweight	Uses common fuels
No moving parts	Develops full thrust on takeoff	Easy to build	
Requires no booster	Has less need for a booster than air-breathing engine	Uses common fuels. Can be solid-fueled	
Can be used in stages or clusters	Can be staged in combination with liquid or solid rockets	Efficient at high speeds and altitudes Supersonic	
Can be stored fully fueled			
Ready to fire at anytime			

TABLE 14-2 *Reaction propulsion disadvantages by type*

Solid-Fuel Rocket	Liquid-Fuel Rocket	Ramjet	Turbojet
High rate of fuel consumption	High rate of fuel consumption	Must be boosted to high speed before it can operate	Limited to low-supersonic speeds, usually less than Mach 3
Short burning time	Short burning time	Limited to operation in oxygen atmosphere	Complicated engine with many moving parts
Comparatively short range unless staged	Comparatively short range unless staged	Speed presently limited to about 3,600 mph	Power limited by stresses on turbine blades
Fragile and sensitive to environmental conditions	Cannot be stored fully fueled for long periods of time		Limited to operation in oxygen atmosphere
	Long checkout procedure at launching		

envelope, and storage requirements. Minimum size and weight may not be the most efficient architecture, and it is often best to employ various types of structures for different sections of the missile to obtain certain design or maintenance advantages. Before addressing missile architecture, consideration must be given to forces experienced by a missile in flight.

Aerodynamic forces are analyzed by resolving the forces into two planar components: lift, which is a force at a right angle to the direction of flow of the medium being traversed, and drag, which is a force parallel to the medium being traversed. The lift and drag forces are considered to act at a point called the center of pressure (C.P.), see figure 14-27.

Lift

Lift is defined as "the net force developed perpendicular to the relative wind." Relative wind is the airflow that a missile experiences as it moves through the air. Lift is created by an airfoil in a relative wind. The airfoil produces a pressure distribution dependent upon its shape and inclination to the passing medium (angle of attack).

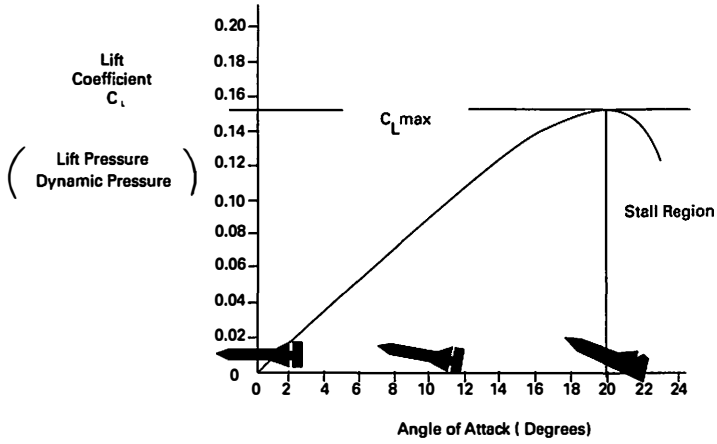


FIGURE 14-28 *Typical lift characteristics*

Lift force is described by the following equation:

$$L = C_L \times S \times \frac{1}{2}\rho v^2 \qquad (14\text{-}8)$$

where:

L = lift as a function of angle of attack α (newtons)
C_L = lift coefficient as a function angle of attack α
S = surface area (m^2)
ρ = density of air ($\approx$1.2 kg/m^3)
v = velocity of the body relative to the airstream (m/s)

Note that the product of $\rho = 1/2\rho v^2$ is called the dynamic pressure (N/m^2).

The lift coefficient is the ratio of the lift pressure to dynamic pressure multiplied by the surface area and is a function only of the shape of the wing and the angle of attack. If the lift coefficient of a conventional wing were plotted versus the angle of attack, the result would appear as in figure 14-28. Since the effects of speed, air density, wing surface area, weight, altitude, and so on, are normalized by using the coefficient form, the true lift capability is demonstrated.

Each angle of attack produces a particular lift coefficient since the angle of attack is the controlling factor in the pressure distribution. The lift coefficient increases with angle of attack up to the maximum, C_L max.

If the angle of attack is increased beyond the $C_L$max angle of attack, the kinetic energy of the boundary layer air molecules is insufficient to keep the air flowing along the contour of the upper surface of the airfoil. This occurs because of an adverse pressure gradient, which allows that air to stagnate. The airflow then separates from the upper surface, causing a drastic loss of lift and increase in "form" drag. This is known as stalled airflow or simply a "stall."

The aerodynamic force of lift on a missile results primarily from the generation of a pressure distribution on the wings and the missile body surface. The entire body of the missile may act as an airfoil and supply lift when it is inclined to the air stream. The missile body at high velocities may provide a major contribution to this force.

The amount of lift created is directly related to the density of the medium and the missile's velocity. Thus, because the size of missile airfoils cannot be increased at high altitudes, loss of lift due to lower air density may occur. This must be compensated for by using an increased angle of attack or a greater velocity.

Drag

Drag is the net aerodynamic force parallel to and in the same direction as the relative wind. Its sources include the pressure distribution and skin friction on the surface. Drag is broken down into parts associated with the various sources of drag. Total drag is made up of the combination of parasite drag and induced drag. Large, blunt bodies in an air stream show a predominance of form drag due to the unbalanced pressure distribution. Streamlined bodies with smooth contours show a predominance of drag due to skin friction. These are both examples of parasitic drag. All bodies that produce lift also produce induced drag. Both types of drag will be discussed in detail below.

The total drag equation is similar to the total lift equation. The basic drag equation is:

$$D = C_D \times S \times \frac{1}{2}\rho v^2 \qquad (14\text{-}9)$$

where:

D = drag as a function of angle of attack α (newtons)
C_D = drag coefficient as a function of angle of attack α
S = surface area (m^2)
ρ = density of air ($\approx$1.2 kg/m^3)
v = velocity of the body relative to the air stream (m/s)

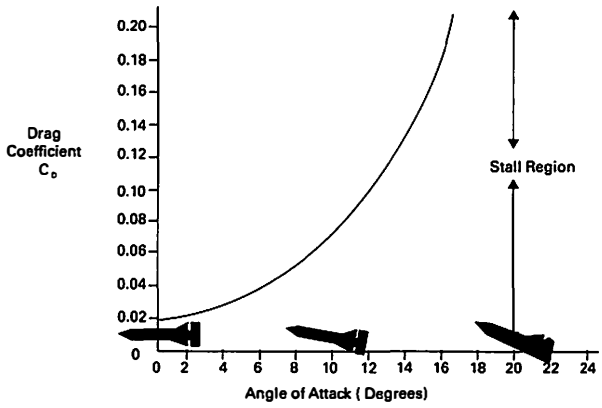


FIGURE 14-29 *Typical drag characteristics*

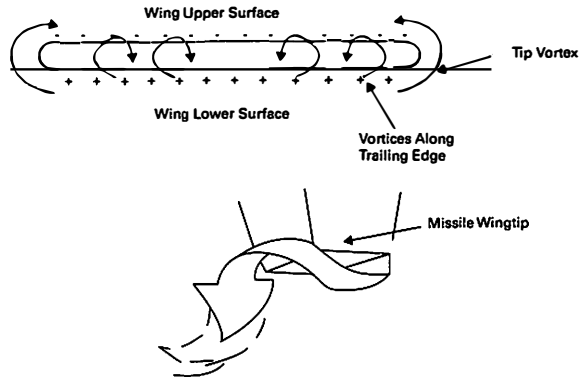


FIGURE 14-30 *Wingtip vortices*

In a fashion similar to other aerodynamic forces, drag forces may be analyzed using a drag coefficient, which is independent of dynamic pressure and surface area. The drag coefficient in this equation is similar to the lift coefficient; it is the ratio of drag pressure to dynamic pressure multiplied by the surface area. The drag coefficient of a conventional missile plotted versus angle of attack is shown in figure 14-29. At low angles of attack, the drag coefficient is low and small changes in angle of attack create only slight changes in the drag coefficient. At higher angles of attack the drag coefficient is much greater and small changes in angle of attack, cause significant changes in drag. As the stall region is entered, a large increase in drag takes place.

As mentioned previously, the total drag force experienced by a missile in flight is a combination of induced drag and parasite drag. These are discussed in detail in the following sections.

Induced drag If a wing is producing lift, a pressure differential will exist between the upper and lower surfaces, that is, for positive lift, the static pressure on the upper surface will be less than that on the lower surface. This pressure differential naturally tries to equalize along the edges of the wing, with the most significant effect created by a spanwise flow from the lower surface to the upper surface at the wing tips. This flow is called a tip vortex and decreases to zero at the wing centerline.

Vortices, shown in figure 14-30, create what is known as downwash, or a downward velocity component (ω) relative to the air stream, see figure 14-31. The total downwash force acts on the wing at a downwash angle (ε), which in effect alters the remote free stream (airflow unaffected by the airfoil) so that the wing now encounters a local relative wind. Lift acts perpendicular to relative wind, so the total effective lift vector now will be inclined aft.

The angle of attack (α) is between the airfoil chord line and the direction of the remote free stream. The effective angle of attack (α_i) is the angle of attack actually experienced by the airfoil and is equal to $\alpha - \varepsilon$. It is this effective angle of attack that determines the lift coefficient for the airfoil, and thus the wing is producing less lift than the actual angle of attack (α) would suggest.

By the definition of lift (force produced perpendicular to the relative wind), the inclination of the local relative wind with respect to the remote free stream causes a component of the "effective" lift vector (L_i) to be parallel to and in the same direction as the remote free stream. This component of lift is known as induced drag (D_i). It is an undesirable, but unavoidable, consequence of any lift-developing device. Figure 14-31 illustrates the generation of induced drag.

In level flight where lift is constant, induced drag varies directly with angle of attack, and inversely with velocity. It can also be shown that there is a direct relationship between lift, induced drag, and the wing aspect ratio.

Wing aspect ratio is means to compare how a wing is built. It is defined as the ratio of the wingspan squared to the area of the wing.

Stated without derivation the lift, induced drag, and aspect ratio relationship is:

$$C_{D_i} = \frac{C_L}{\pi \cdot e \cdot AR} \qquad (14\text{-}10)$$

where:

C_{D_i} = induced drag coefficient
C_L = lift coefficient
AR = wing aspect ratio
e = plane form efficiently (essentially the shape of the wing with ranges between 60%–95%)

This detailed explanation of induced drag is provided to underscore the fact that any airfoil creating lift is experiencing this "opposing" force to some degree. It is

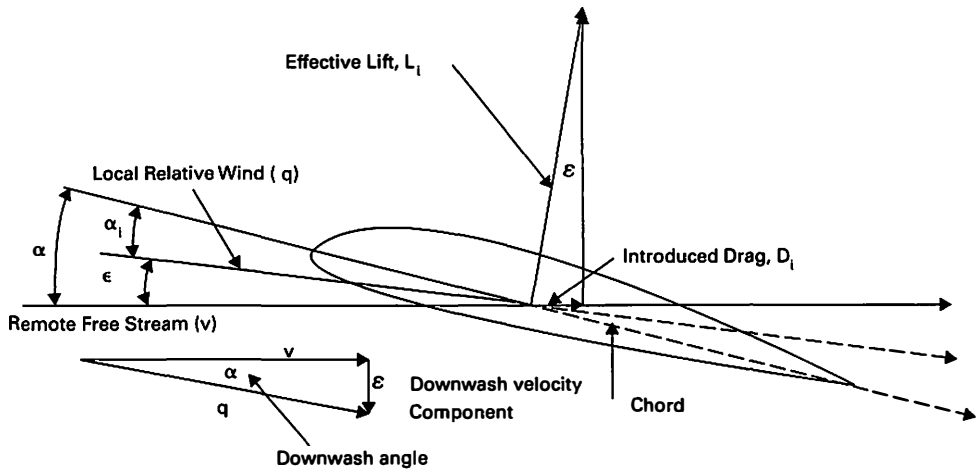


FIGURE 14-31 *Induced drag*

essentially the "cost of flying." Equation 14-10 facilitates an understanding and appreciation of induced drag.

Flight at high coefficient of lift conditions, such as low speed or maneuvering flight, will create high induced drag whereas high speed, low C_L flight will create low induced drag. Thus, a high speed missile in relatively straight and level flight will have a very small amount of induced drag.

The effect of aspect ratio is significant in that the lower the aspect ratio the higher the induced drag. This is because airfoils with low aspect ratios must operate at high angles of attack to achieve maximum lift. Induced drag and the aerodynamic effect (downwash angle) that causes it may be reduced by designing wings with higher aspect ratio or by adding devices such as tip plates or winglets to an existing wing in order to reduce spanwise airflow.

Parasitic drag The sum of all drag not associated with the production of lift is termed parasitic drag. Parasitic drag is a force that opposes motion of the missile, due to displacement of the air in front of it and due to friction from the flow of air over all its surfaces. Typically parasitic drag is broken down into form drag, friction drag, and interference drag.

Form drag, also known as pressure or profile drag, is caused by the airflow separation from a surface and the wake that is caused by that separation. It is primarily dependent upon the shape of the object.

Friction drag is associated with the air that directly contacts the missile surfaces. Rough missile surfaces produce friction as air flows over them. Reduction of skin friction also allows smooth airflow close to the surface of the missile. Smooth boundary layer airflow is called laminar flow. Parasitic drag associated by friction and interference is minimized for that particular configuration that produces the most laminar flow.

Interference drag is generated by the mixing of airflow streamlines. Abrupt meeting of streamlines creates greater interference drag, while proper fairing to enable streamlines to meet gradually minimizes it. If the surface of the missile is irregular, turbulence in the flow of air around the missile will increase and counteract the effect of streamlining. By minimizing these irregularities, the layers of air close to the surface of the missile will flow more smoothly.

Although parasitic drag is not directly associated with the production of lift, it does vary with angle of attack and lift. Parasitic drag is directly proportional to the velocity squared, thus streamlining of the missile becomes paramount as velocity increases.

The variation of the parasitic drag coefficient, C_{Dp}, with lift coefficient, C_L, is shown for a typical missile in figure 14-32. The minimum parasitic drag coefficient, $C_D$min usually occurs at or near zero lift and increases above this point in a smooth curve.

Total drag The induced drag coefficient is also shown in figure 14-32 for purposes of comparison, and because the

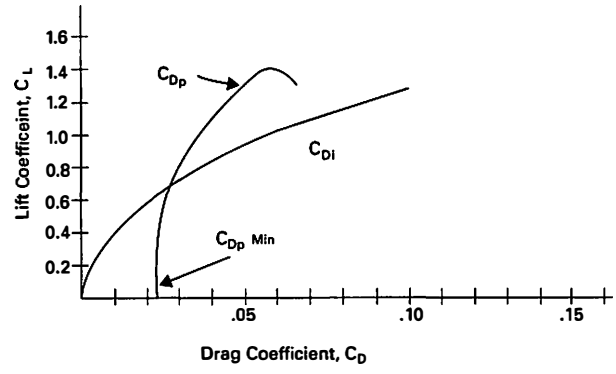


FIGURE 14-32 *Typical parasitic and induced drag curves*

total drag of the missile is a sum of the parasitic and induced drag.

Figure 14-32 shows that both C_{D_i} and C_{D_p} vary with lift coefficient, and therefore, with the angle of attack. Note that C_{D_p} is nearly constant at low angles of attack, compared to C_{D_i}. Because missiles usually incorporate high thrust motors, high velocities are attained. These high velocities are the major contributor to the production of lift, and angles of attack are normally very low (<5°) and thus the assumption is made that C_{D_p} is, in fact, a constant.

Referring to equation 14-9 it would seem to be a simple matter to solve for the parasitic drag for a missile in balanced flight. However, with the exception of a cruise missile, a missile is never in balanced flight. During powered flight a missile accelerates at a non-linear rate, and after its motor burns out, while coasting, it decelerates at a constant rate. Thus the velocity term of the drag equation is constantly changing, thereby continuously changing the amount of drag.

To accurately predict a missile's flight path and achieve a fire control solution, parasitic drag must be taken into account when calculating acceleration, velocity and distance traveled. For a missile with constantly changing acceleration, this requires the solution of a very messy non-linear differential equation during powered flight and a linear differential equation for coasting flight. This would be an altogether disagreeable task if it weren't for the digital computer.

Stability

Although the weight of the missile is considered to act at the center of gravity (C.G.), in figure 14-27 the lift and drag forces are considered to act at the center of pressure (C.P.). Lift and drag force vectors intersect at this point, and if the center of gravity and the center of pressure are not collocated, as is often the case, the resulting moment of forces acting at these points may cause unstable flight conditions. Therefore, the location of the center of pressure, determined by the configuration of the missile's structure, is an important factor in missile stability.

In addition to the moments that tend to produce rotation about the center of gravity (roll, pitch, and yaw), there are moments applied about the hinges or axes of the missile control surfaces. When force vectors representing these pressures have their origin at the center of gravity, it is then possible to describe rotation of the missile in terms of angular displacements about these axes.

When all forces acting on the missile are resolved into component forces along the three axes, then the resultant motion of the missile with respect to its center of gravity can be completely specified.

A missile in flight encounters various forces that may be characterized as disturbing forces or moments, inasmuch as they tend to cause undesired deviations from the operational flight path. They may be random (such as wind gusts) or systematic (such as misalignment of thrust components). When a missile is designstable, it will return to its former position after being disturbed by an outside force. Guided missiles have the great advantage over unguided missiles in that guided missiles are able to correct, to some degree, both random and systematic disturbances that may occur during flight.

Control Surface Configuration

In guided missiles it is desirable to have a method of altering the direction of motion. By including airfoils with variable angles of attack, forces can be applied to alter the orientation of the missile in a plane at right angles to the plane of the airfoil, just as the rudder of a ship alters its angle to the flow of water.

In both cases, it is this change in angle of attack of the body of the vehicle that results in a turn. Variation of the location of control surfaces and other airfoils can significantly change the performance characteristics of the missile. Control surfaces can be classified as canard, wing, tail, or a combination of these surfaces.

Canard control is characterized by small control surfaces that are well forward on the body, whereas the main lifting surface is well aft; see figure 14-33. These control surfaces are deflected in a positive manner; that is, the leading edge is raised to provide for a positive attack angle.

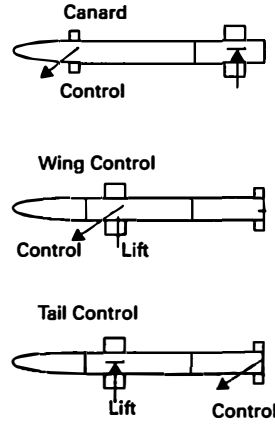


FIGURE 14-33 *Control surface placement*

To apply sufficient force, the canards must be positioned at a large angle of attack, causing large loads on hinges and missile structure. Therefore, large amounts of power are required to position these surfaces rapidly in order to reduce the total time that the force and additional drag are experienced. Canards require relatively large fixed surfaces aft on the missile airframe to improve stability.

Wing control includes control surfaces near the center of the airframe, which are also the main lifting surfaces. The entire lift surface can be controlled to increase or decrease lift in response to control signals. This produces rapid missile maneuvers without requiring large inclinations to the airstream, which results in minimum drag and minimum change in missile attitude or orientation in space.

Maneuvers are initiated at the instant the wings are deflected and are not dependent upon an increase in body angle. With this method, delay in change or missile orientation may be detrimental to some guidance systems because the missile seeker may not be able to maintain contact with the target during maneuvers.

Tail control surfaces are located at the rear of the body. Lift, supplied by fixed airfoils at the midsection, and deflection of the tail control surfaces are used to alter the missile angle of attack. With this configuration, control surface deflections are in a direction opposite to that of the angle of attack. Deflections are minimized, resulting in a low control surface and low body-bending loads, as well as in little force on control surface pivots or hinges. Since the main lifting airfoils are not deflected, wing tail interference effects are minimized.

Because of the tail control surfaces' distance from the center of gravity (long lever arm), control surface size can be reduced, resulting in minimum drag. Control mechanism design problems, however, are compounded by the small volume available and the severe heat conditions that may be encountered when the control surfaces are located aft in the vicinity of the propulsion system exhaust nozzle.

Missile Components

The components of a missile are located in five major sections; see figure 14-34:

1. Guidance system
2. Warhead section
3. Autopilot
4. Control system
5. Propulsion system

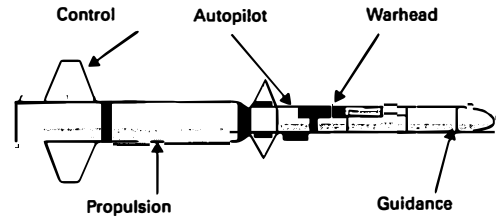


FIGURE 14-34 *Functional systems of a guided missile*

Guidance System

The guidance system for a homing missile consists of an antenna assembly or electro-optical device protected by an aerodynamic cover. Additionally, this section contains electronic components that analyze signals from the target and compute orders for use by the autopilot. The sensor employed is usually a gimbal-mounted automatic tracking sensor (except the interferometer method) that tracks the target line-of-sight (LOS) and sends signals about the target's movement to the guidance electronics.

Warhead

The warhead consists of the fuze assembly, booster, and primary explosive warhead. The fuze assembly usually contains the safety and arming device and commonly has both a contact fuze and a proximity fuze. The contact fuze provides redundancy to the proximity fuze. The proximity fuze, also known as a target detection device (TDD), receives a signal from the guidance section to activate just prior to intercept, minimizing vulnerability to countermeasures. The safety and arming device prevents arming of the warhead until the missile is a safe distance from the firing platform.

Autopilot

The autopilot is a set of electronic instruments and electrical devices that control the electric actuators (motors) of aerodynamic control surfaces (fins). In absence of signals from the guidance computer, the autopilot maintains the correct missile attitude and maintains the missile flight in a straight line. Called-for-acceleration signals from the guidance computer will cause the autopilot to command corresponding changes in flight path while maintaining stabilized missile flight.

Propulsion System

Any of the methods of propulsion previously described may be used as long as the missile has sufficient speed advantage over the target to intercept it. The propulsion system must rapidly accelerate the missile to flying speed to allow a short minimum range and to achieve sufficient velocity to counter target maneuvers.

Powered flight may occur for most of the operational range of the weapon or only at the beginning (boost-glide). Boost-glide weapons are limited in their ability to engage long-range targets that have a significant altitude difference or that perform rapid maneuvers.

Control System

The steering or control unit may be located forward, in the midsection, or aft on the missile, depending on where the control surfaces are located. The movement of control surfaces may be electrical, hydraulic, or pneumatic. Some weapons are limited in allowable locations for the control actuators because of size limitations or difficulty in passing signals from the autopilot to remote points on the airframe.

Architecture of Gun Ammunition

Gun projectiles are similar in function to missiles, but they differ in their means of propulsion and stabilization. Their external shape is designed to obtain maximum stability while experiencing minimum drag. The internal structure of the projectile is dictated by the type and strength of the target, gun pressure characteristics, and the type of trajectory desired.

The distribution of weight is a matter of considerable importance in that the center of gravity should be on the longitudinal axis and forward of the center of pressure to ensure maximum stability. High standards of accuracy are required in gun projectile manufacture to ensure uniform flight characteristics and therefore predictable hit patterns.

Parts of a Projectile

Small arms and machine-gun projectiles are made of solid metal; however, projectiles of 20-millimeter guns and larger have many components. The external features of projectiles are depicted in figure 14-35. The form of the forward end of the projectile is an ogival curve (generated by revolving an arc of a circle about a chord) that is aerodynamically efficient.

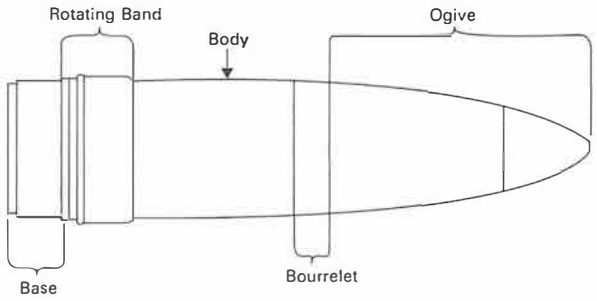


FIGURE 14-35 *Details of a gun projectile*

Behind the ogive, the projectile's body is cylindrical with the exception of the bourrelet, which is slightly larger than the diameter of the body in order to reduce the surface area and thus the friction of the projectile contacting the gun bore.

Near the after end of the projectile is the rotating band, which is actually larger than the gun bore diameter to allow for engagement of the rifling grooves and to ensure a gas-tight seal in the bore while supporting the aft end of the projectile. Aft of the rotating band, the cylindrical shape may continue to the base of the projectile, or it may be tapered to a "boat tail." A square aft section is more efficient in sealing propellant gases and converting pressure to kinetic energy; however, the boat tail provides better aerodynamic characteristics.

Some projectiles such as that in figure 14-36 incorporate a plastic rotating band and rear section that discards in flight, leaving a boat tail rear section and reducing air friction due to the rough surface of the rotating band. This projectile has less drag and therefore more range than similar projectiles without this capability.

Aerodynamic Drag

All drag forces act at the center of pressure and are in the opposite direction of the motion of the projectile; see

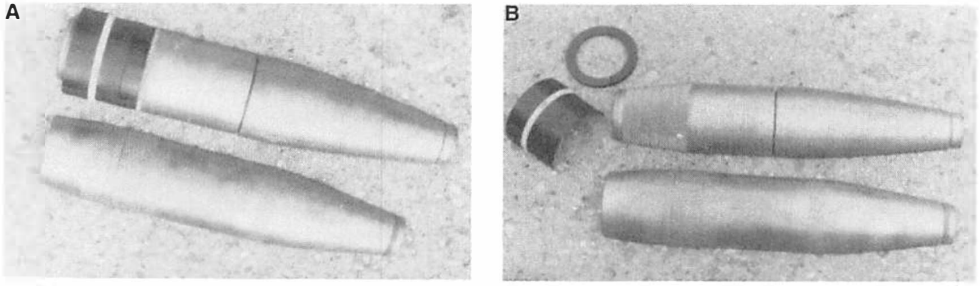


FIGURE 14-36 *A five-inch projectile showing the difference between a discarding aft section (the top projectile) and a conventional projectile underneath it. Note the more pronounced taper of the rear section of the new projectile once the rotating band is discarded, B.*

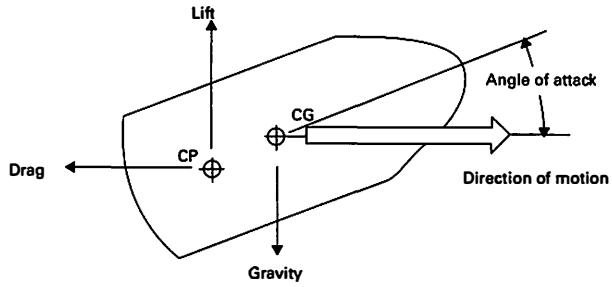


FIGURE 14-37 *Aerodynamic forces on a gun projectile*

figure 14-37. Again, the drag force is calculated using equation 14-9.

The configuration of a moving projectile determines the manner in which air will behave as it flows over the projectile's surface. A tapered or pointed projectile encounters less resistance in its movement through air than a blunt-nosed projectile, and a projectile with a tapered base allows air to flow by it more readily than one with a square base.

Figure 14-38, a plot of the drag coefficient against Mach number for the projectile shape illustrated, indicates an increase in the value of the drag coefficient as the projectile approaches the speed of sound. The sudden increase in drag in this region is due to the fact that as a projectile moves through the air, the air tends to compress and pile up in front of the projectile, and can flow smoothly past it only at velocities less than Mach 1.

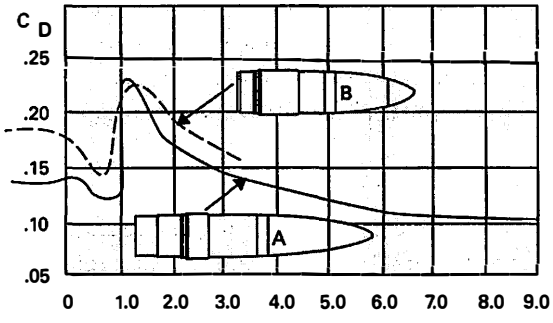


FIGURE 14-38 *Drag coefficient versus Mach number*

As the projectile velocity approaches the speed of sound, the air can no longer flow past the surface of the projectile, forcing the projectile to split the airstream, which produces shock waves and increases drag coefficients. At speeds greater than that of sound, the entire character of the airflow is changed. At lower velocities, a projectile is retarded primarily by the skin friction between the airstream and the projectile surface. As the velocity is increased, the airstream is unable to close in behind the base, and a decided turbulence appears behind the projectile (a wake). The projectile must then overcome drag coefficients from both skin friction and wake.

A further increase in velocity beyond the speed of sound will introduce resistance in the form of shock waves. Therefore, a projectile traveling at supersonic speed encounters retardation, insofar as velocity is concerned, which is the combined effect of skin friction, wake, and shock waves.

Aerodynamic Stabilization

Two methods are employed to stabilize projectiles in flight: fin stabilization and spin stabilization. Most projectiles are stabilized by a spin imparted by the rifling (helical or spiral) in the bore of the weapon. The twist of the rifling determines the projectile's rate of spin. The twist may be "uniform," generally about one in fifteen or twenty times the bore diameter, or "increasing," as in the 40-millimeter gun, causing the twist to sharpen as it nears the muzzle.

Regardless of how the spin is imparted, the spinning causes gyroscopic forces to act on the projectile for stabilization in flight. Projectiles launched by guns without rifling may utilize fins to control flight.

There are three general classes of projectiles:

1. Penetrating
2. Fragmenting
3. Special-purpose

Penetrating Projectiles

Penetrating projectiles, sometimes referred to as armor-piercing (AP), employ kinetic energy to penetrate heavy armor. The bursting charge employed in the projectile must be insensitive enough to permit penetration without premature detonation. Figure 14-39 shows the architecture of an AP projectile having thick steel walls and a hardened nose cap to penetrate armor or concrete, thus reducing the amount of space for explosives.

Because efficient penetrating shapes are not very aerodynamic, an ogival-shaped thin steel windshield is included to reduce drag. These are the heaviest projectiles

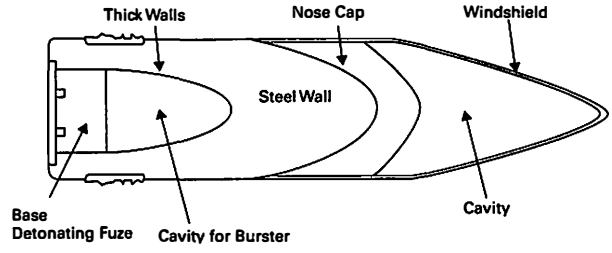


FIGURE 14-39 *Armor-piercing projectile*

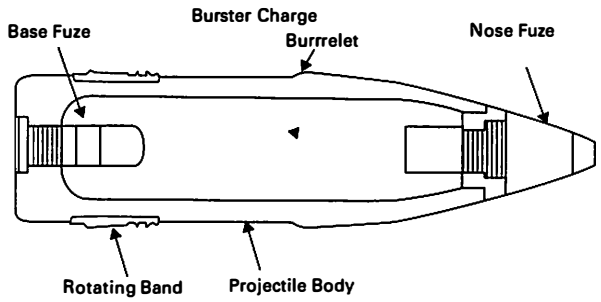


FIGURE 14-40 *Multipurpose fragmentation round*

and therefore have the lowest initial velocity, but their velocity does not decay as rapidly as other projectiles' velocity due to the additional mass.

Fragmenting Projectiles

Fragmenting projectiles (see figure 14-40) inflict damage by separating into high-velocity fragments upon detonation. They have relatively thin walls that may be scored and have a large cavity for the bursting charge. Because of the use of multipurpose fuzes, this type of round is used for both air defense and shore bombardment.

Special-Purpose Projectiles

Special-purpose projectiles are designed to support the destruction of targets, such as by providing illumination or screening smoke. If the payload includes an explosive, its small charge is designed to expel the contents of the projectile.

Guided Projectiles

The idea of guided projectiles came from the outstanding performance of the laser-guided bombs used in Vietnam. Guided projectiles employ a miniature shock-resistant semi-active laser seeker or GPS receiver in conjunction with small folding control surfaces for terminal guidance and may or may not include a rocket assist (RAP) motor.

The Navy, in conjunction with the Army and Marine Corps, developed a family of 5-inch, 155-millimeter, and 8-inch projectiles that in some applications provide greatly increased hit probability at much less cost than a guided missile.

Gun Propellant Configuration

Case guns are those that employ propellant encased in a metal shell, whereas older gun systems employed propellant charges packed in silk bags. The use of bags was confined to very large guns where the total propellant powder necessary to attain the required initial projectile's velocity was too great in weight and volume to be placed into a single rigid container.

For ease of distinguishing, ammunition and powder charges used today are classified as either fixed or semifixed. Fixed ammunition, often referred to as a cartridge, has the propellant charge and the round joined together; this type is used for small bore guns. For ammunition generally larger than 76 millimeters, semifixed propellant charges, where the charge is separate from the round and is joined prior to loading in the gun barrel, are used; see figure 14-41.

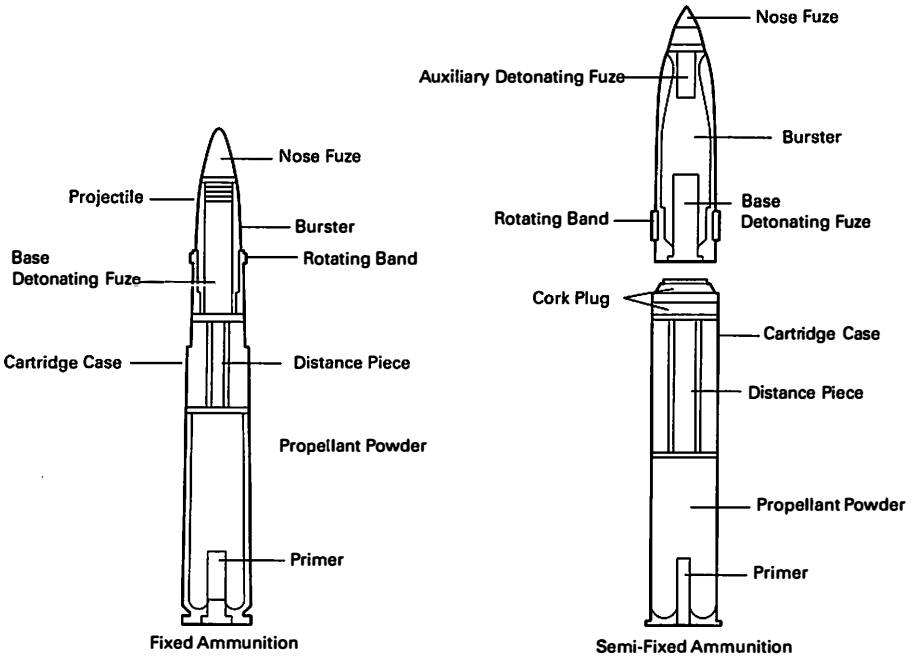


FIGURE 14-41 *Cross section of gun propellant configurations*

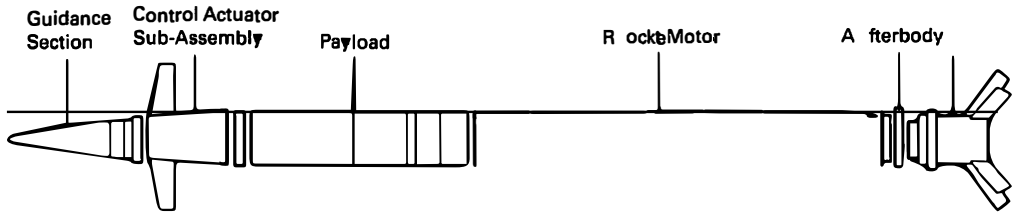


FIGURE 14-42 *A long-range land attack projectile for the advanced gun system; it has a 155 millimeter round with an expected range of 100 nautical miles.*

Special Cases

There is an ever-increasing need for projectiles that travel at faster velocities. The advantages of a faster projectile are that more kinetic energy can be imparted to a target, the range of the weapon system increases, and the fire control problem becomes easier. Special gun and projectile designs are constantly being developed that employ variations in or higher-level improvements of previously used gun-manufacturing techniques. To increase the velocity of the projectile, either the muzzle velocity must be increased or the round itself must change configuration upon exiting the gun barrel.

The basic principle of muzzle velocity increase is to increase the area under the projectile pressure travel curve. For an optimum gun, this requires approaching a pressure travel curve that is nearly as flat as possible. The following are practical examples of various means employed to achieve higher gun projectile muzzle velocities.

Lightweight projectiles By firing lightweight projectiles from conventional guns, the muzzle velocity for standard projectiles can be increased from 2,700 feet per second to about 4,000 feet per second. This type of round, which is greatly affected by air resistance, is useful only at short ranges.

Subcaliber projectiles By using a projectile of smaller caliber than the gun, the projectile weight and the resistance encountered during flight are reduced. In order to seat the projectile firmly in the gun, the projectile is fitted with a lightweight bushing, called a sabot, which falls away from the projectile after it leaves the muzzle.

High-strength, extra-long guns Conventional projectiles fired from high-strength or extra-long guns can attain a velocity limited only by the amount of pressure the gun can withstand and by the pressure that can be developed with the available type of powder. Such guns are heavier and require considerably more space aboard a ship than conventional guns and, because of increased erosion, wear out more quickly than guns designed to fire projectiles with conventional velocities.

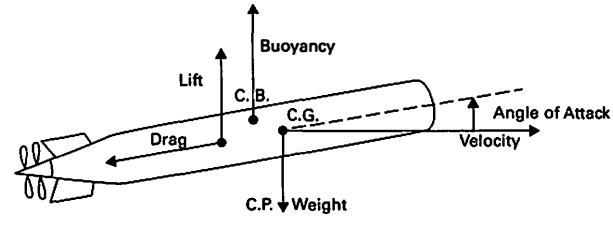
Rocket assisted projectiles Another method of achieving higher projectile velocities is the employment of a moving propellant. The projectile consists of a solid-propellant rocket motor with a delayed ignition element, an explosive-filled warhead, and a guidance package with control surfaces; see figure 14-42. When combined with the Global Positioning System, microprocessors, and inertial guidance, this type of round increases the accuracy and range compared to conventional gun systems.

A problem that currently exists with this system, however, is that the shock of firing from the gun puts great stress—sometimes on the order of 12,000 g's—on the guidance package, rendering the round ineffective.

Architecture of Torpedoes

The technology acquired in the development of advanced guided missiles is generally applicable to the torpedo. There is, in fact, a direct analogy between the homing missile and the homing torpedo that allows a similar approach in their discussion by substituting the theories of hydrodynamics for those of aerodynamics.

Air and water are fluids that display some of the same properties when allowances are made for differences in density and mass, and the general lack of compressibility of water. The force diagram for the torpedo as depicted in figure 14-43 is similar to that for the missile, except that buoyancy is included.



C.P. = Center of Pressure
C.B. = Center of Buoyancy
C.G. = Center of Gravity

FIGURE 14-43 *Torpedo force diagram*

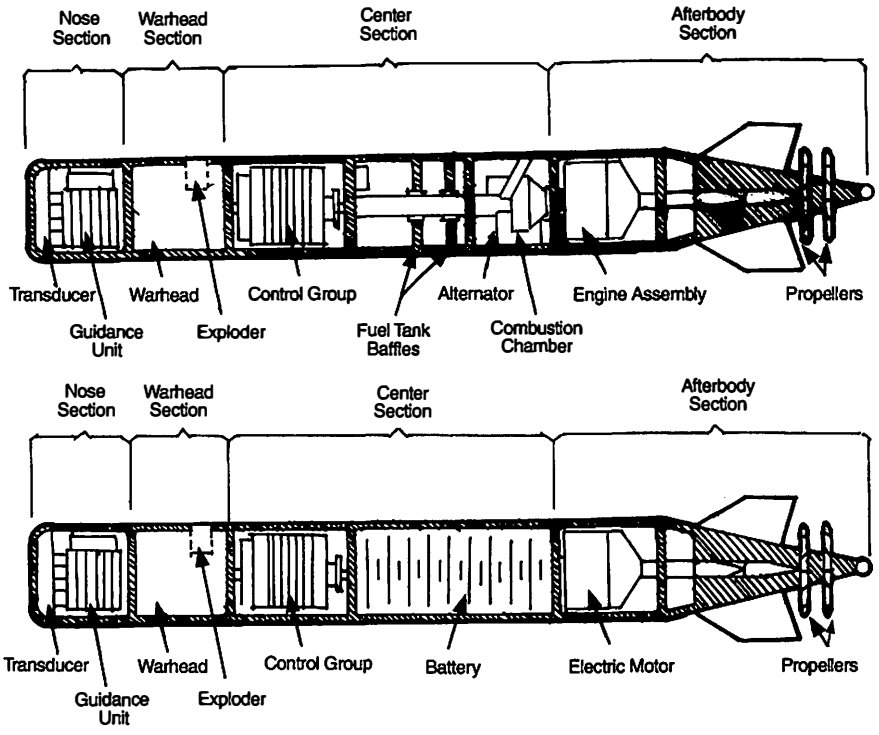


FIGURE 14-44 *Diagrams of an external combustion engine torpedo (above) and an electric torpedo (below)*

To understand how the torpedo operates, a study of the following functional components is required:

1. Propulsion system
2. Control and guidance system
3. Warhead and fuze

Some of the major components of liquid fuel and electric torpedoes are shown in figure 14-44.

Propulsion System

Components of the propulsion system—either an internal, or external combustion-type engine or an electrical battery-powered engine—are physically located in the torpedo center and afterbody sections. Upon the torpedo's water entry the propulsion system is activated. If it is a liquid-fueled torpedo, the fuel flows to the combustion chamber. Hot gases produced from burning fuel drive the engine or turn an alternator to provide electrical power that propels the torpedo via two counterrotating propellers.

This fuel is a monopropellant—that is, it contains the oxygen necessary for its combustion and, like most liquid propellants, is very toxic, combustible, and corrosive.

In some torpedoes a solid fuel similar to that described for rockets is used to supply hot gas to the engine. If it is an electrical-type torpedo, a set of batteries or a fuel cell provides the power to the electrical motor.

Because propulsion power requirements increase as a function of depth, it is necessary to increase engine output as depth increases to maintain exhaust flow due to higher ambient pressure. To accomplish this, a hydrostatic device on the fuel pump increases the quantity of fuel supplied to the combustion chamber, which in turn increases hot gas production and engine output. This increased fuel consumption causes a loss of torpedo running time and range at greater depths.

Control and Guidance System

Advanced torpedoes have very sophisticated control and computing systems that give these weapons, once launched, autonomous control capabilities.

Torpedoes can employ either active or passive homing and in some cases can receive command guidance via a trailing wire. In active homing, the torpedo transmits sound and listens for a returning echo. In passive homing, the torpedo detects noise radiated from a target. If no signal is received, the torpedo can shift automatically to active search.

After reaching its operational depth, the torpedo can search in a circular or other preprogrammed path. The type of search conducted depends on the launch platform, torpedo type, and target category.

The torpedo detects targets by receiving acoustic energy that can be evaluated through the use of transducers to

determine the validity of the contact and the target's bearing and range. The transducer array can consist of many elements that operate by employing the electronic scanning principles discussed in chapter 11 to create a directional beam and change its direction.

A narrow vertical beam is used in shallow water and during the attack phase to minimize surface and bottom reverberation, whereas a broad vertical beam is used in deep water for better target acquisition. The computer or automatic pilot decides which beamwidth is appropriate.

Whenever the torpedo is searching in the active mode, the transmitter is controlled by the computer and, when directed, generates a pulse of energy that is sent to the transducer through a series of amplifiers. Computer inputs control the characteristics of the pulse such as its power, duration, and interval.

Basic functions of the receiver are to receive target information from the transducer, to determine if the contact is a valid target, and to supply valid target signals to the computer. Self-noise and reverberation are detrimental to target detection, but can be suppressed by special circuits that distinguish a target's noise from background noise.

The computer provides control signals and a time base for the transmitter, receiver, and automatic pilot. When the torpedo reaches its programmed search depth, the automatic pilot sends a signal, enabling the computer to function. At the same time, it informs the computer whether homing is to be active or passive, and the type of search pattern to use.

Upon receipt of a valid signal, the torpedo determines range by calculating the time interval between transmission and reception. Next, the torpedo sends a range gate back to the receiver that inhibits reception of any echo not at approximately the same range as the previous valid echo. After transmission of the range gate, the yaw angle signal from the receiver is converted to right or left commands to the autopilot.

This command starts the torpedo steering toward the target bearing. The computer updates its command signals after every echo in order to guide the torpedo to the target. At appropriate times during this attack phase, information is sent to the transmitter to change its power, vertical beamwidth, and beam axis.

The automatic pilot receives torpedo programming information before the launch, transmits part of this information to the other components in the control and guidance system, controls certain depth functions, and combines command signals with existing conditions to control a steering actuator. A course gyro is contained in the autopilot, which senses the torpedo heading in relation to a programmed heading set at launch.

The autopilot also contains the yaw, pitch, and roll-integrating gyros. Steering commands for pitch and yaw, whether generated by the computer or autopilot itself, are combined with the rate gyro's outputs so that correct command signals are sent to the actuator controller. That is, the autopilot combines desired results with existing conditions of motion and orders the fin controllers to reposition the fins.

The steering-control actuator contains three motors, which are controlled by signals from the power unit. The outputs of the motors move four fins that control the movement of the torpedo; the upper and lower rudder fins are operated individually by separate actuators to control yaw and roll; the port and starboard elevator fins are operated simultaneously by a common actuator to control pitch.

Warhead Section

Torpedo warheads are either pure blast devices or special warheads designed to penetrate well-protected targets. Because fuzes are covered in chapter 16 and warhead effects in chapter 18, they will not be further addressed here.

Architecture of Naval Mines

A naval mine is a weapon that is laid or planted to be detonated some time in the future, either remotely or by the actions of a target. Naval mines are often employed in groups called minefields, which are arranged in a particular pattern to provide a desired density and area coverage.

It has been said that a mine is a piece of ordnance whereas a minefield is a weapon; as such, minefields continue to be an extremely effective means of denying the use of a particular area to ships and submarines. Naval mines are weapons that are not required to do anything to achieve their purpose except to pose a threat, real or imagined.

Classification

Naval mines are classified by their method of delivery, placement, or activation (covered in chapter 16 on fuzing). The most common delivery method is by aircraft when large minefields are needed quickly but precise placement of the weapon is not needed. Ships and submarines can place mines with greater accuracy, and submarines have the advantage of being able to place mines covertly. It is important to note that mines can be placed by a variety of craft and techniques and that unconventional covert methods are very effective and are becoming more common.

Naval mines are designed to be buoyant or nonbuoyant depending on their intended final position. Regardless of buoyancy considerations, all naval mines must be

made to withstand the environment's corrosive effects in order to maintain watertight integrity. The following is a list of the common designations of mines based upon their placement position; see figure 14-45.

Moored mines Buoyant bodies that are anchored to the bottom are called moored mines. A specific length of mooring cable is used to allow the mine to float at a predetermined depth. Good-quality hydrographic surveys and knowledge of tidal ranges are critical to the placement of this type of mine.

Bottom mines This type of mine is designed to rest on the seafloor and is usually deployed in relatively shallow water.

Oscillating mines Sometimes called drifting mines, these are neutrally buoyant mines that are designed to float untethered at a predetermined depth. These mines drift with the prevailing currents and consequently are very dangerous to all seagoing traffic. This type of mine is not used by the U.S. Navy.

Mobile mines A type of mine that propels itself to a position and then becomes a bottom-type mine is called a mobile mine. This type of mine is safer for the laying platform because it allows the planting of mines at a standoff distance from a well-defended area. This method also provides an element of surprise by concealing-from the enemy the presence of a minefield.

Propelled warhead mines This type of mine first starts as a bottom or moored mine and then becomes a movable warhead. When a target is detected, a self-propelled warhead section detaches from the mine case or anchor to close the target. This type of mine can be a torpedo or an underwater rocket-type weapon. The U.S. Navy's Captor mine is an example of this type. It releases an MK-46 torpedo at a target within its detection range.

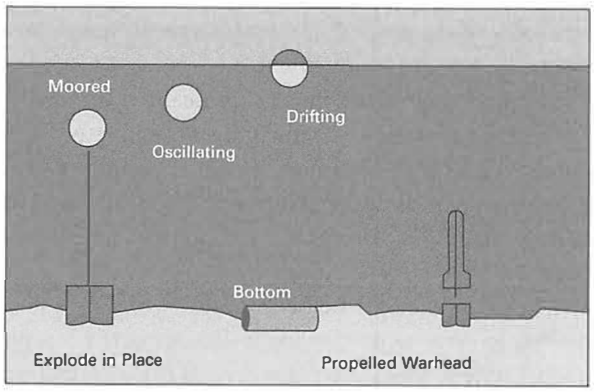


FIGURE 14-45 *Basic mine positioning*

Mine Components

Mine components vary widely depending on the type of mine and the specific purpose of employment. However, all mines have a case and a warhead section. Other mine components, depending upon application, provide for mine control, countermeasure resistance, and sterilization.

The mine case is the main element of the mine. It contains or carries the explosive system and all those components by which the mine arms, detects a target, actuates, and, in some cases, scuttles itself or sterilizes under specific conditions.

In the moored mine, the case and the anchor are married together until the mine is laid, at which time they separate; the anchor rests on the sea bottom and the mine case assumes a position above the anchor to which it is connected by its mooring cable. Figures 14-46 and 14-47 illustrate typical moored mines and their major components.

Since the bottom mine has no anchor, the mine case must be of sufficient weight to be stable on the bottom. The large amount of explosives that may be carried is characteristic of the bottom mine, which can cause great damage if properly positioned.

The type of laying agent plays an important part in regulating the exterior shape and size of mines. Air-laid mines are dropped from bomb stations on aircraft and, therefore, tend to be cylindrical in shape. Submarine-laid mines can be placed through the torpedo tubes or from belts placed outside the hull. A torpedo tube-laid mine tends to be long and cylindrical, whereas mines from belts can be a variety of shapes.

Mines laid by surface craft have little or no restrictions on their shape; these types tend to be spheres, or modified spheres, positioned on top of box-shaped anchors that are fitted with wheels.

The mine case is generally made of sheet steel, but aluminum, spun glass, or various plastics may also be used in some mine types. Mine casing openings are carefully sealed with suitable gaskets to maintain watertight integrity and resist water pressure at great depths.

The warhead section consists of the explosive components common to all explosive-type weapons, namely, the detonator, booster, and main charge. How the warhead functions and detects the target is covered in detail in chapters 16, 17, and 18.

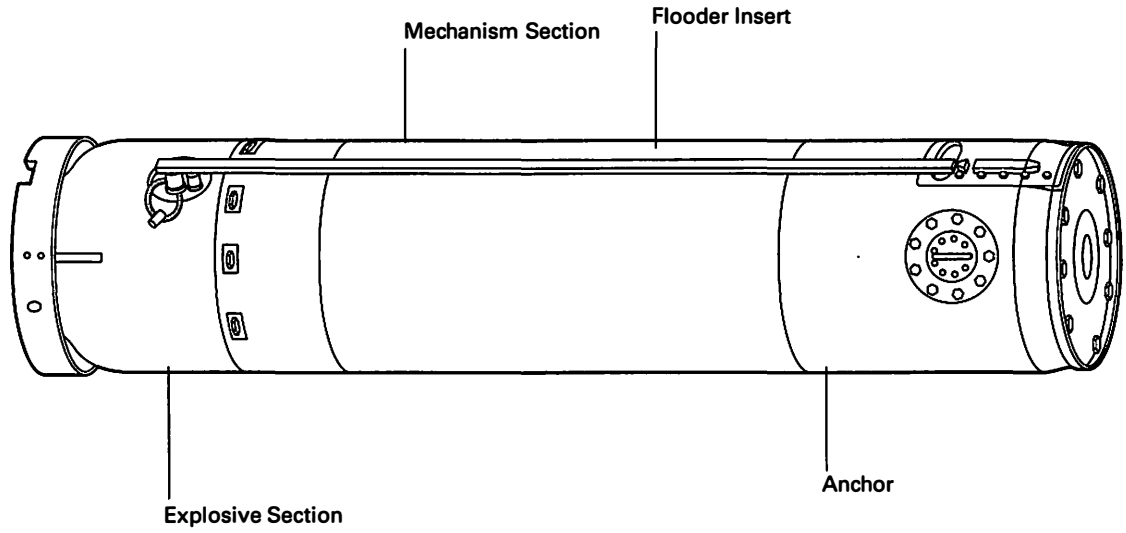


FIGURE 14-46 *Submarine-launched naval moored mine*

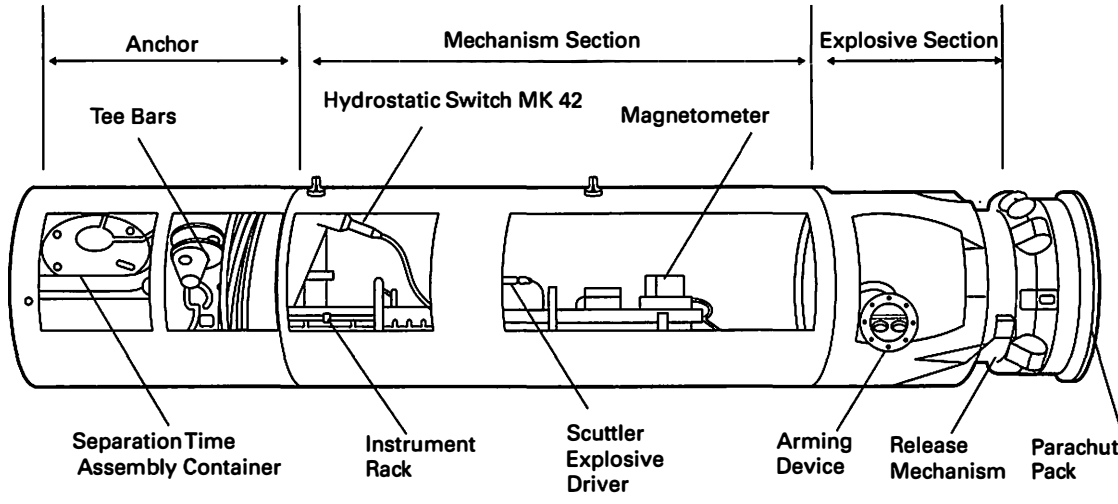


FIGURE 14-47 *Typical air laid moored mine*

Summary

The purpose of a launching system is to place a weapon into a flight path as rapidly as possible. Launching systems must perform with speed and reliability, while displaying weapon-system compatibility and safety. The basic functions of the launching system are weapon storage, transfer, loading, control, and launching.

The three basic types of launchers are gravity, impulse, and reaction. The lightest and least-complex type is the gravity launcher, whereas the impulse launcher is characteristically heavy and bulky. The requirement for impulse launchers to contain the propellant gases and to dissipate the energy of the propellant through a recoil system contributes to their size requirement.

Reaction launchers have significantly minimized the size and weight disadvantages associated with impulse launchers. However, weapons employed with reaction launchers are characteristically more complex than the weapons associated with gravity and impulse launchers.

Launchers, once assigned to the weapon system, respond to error signals, called orders, from the computational system; these orders are derived from present target position as established by the sensor system.

Once the weapon has been launched, the propulsion system propels the weapon to the target. The three main types of propulsion systems are impulse, reaction, and gravity. Impulse propulsion includes the study of interior ballistics and the propellants used in today's guns and impulse-propelled weapons.

Reaction propulsion deals with systems that internally carry their own propulsion. Rocket engines and thermal jet (air-breathing) engines are two types of reaction propulsion systems. Within the rocket engine category are both liquid and solid-fueled engines, and within the thermal jet engine category fall the turbojet and ramjet engines.

Finally, gravity propulsion is the propulsion system used for dropping bombs and relies on the principle of converting potential energy to kinetic energy.

Weapon architecture is concerned with placing the components necessary for proper operation in such an order and configuration that optimum performance is achieved. For missile systems, aerodynamics and sensing systems drive the structure. For gun systems, payload and balance are most important. Torpedoes operating autonomously require very accurate guidance, control, and sensing systems, whereas naval mines' primary design constraint is that they must be able to withstand the environment for long periods of time.

---

[i] DARPA/ONR/NASA Langley Research Center

15 Guidance Systems

Introduction

The term *missile* in the post-World War II era has generally been used synonymously with *guided missile*, due to the wide impact of guided-missile technology upon the weapon field. In the unguided case, initial conditions—that is, train, elevation, powder charge, barrel temperature, and so on—and exterior ballistics are parameters that, along with normal distribution, affect the "fall of shot."

As advances in technology permitted, which was paralleled by increasing threat complexity, development of guided weapons made possible a significant increase in terminal accuracy of warheads. The application of automatic controls is prevalent in broad regions of weapon technology, including:

- Underwater homing torpedoes
- Surface-to-surface aerodynamic guided missiles
- Intercontinental ballistic missiles
- Air-to-air guided missiles
- Surface-to-air guided missiles
- Air-to-surface guided bombs
- Guided projectiles

This chapter will treat primarily the aerodynamic guided missile; however, other types of weapons will be discussed as applicable. Various aerodynamic missile flight paths will be introduced and a simplified analysis that includes the relative advantages and disadvantages of each will be given.

Fundamentals of Guidance Systems

Purpose and Function

Every guidance system consists of an attitude control system and a flight path control system. The attitude control system functions to maintain the weapon in the desired angle of attack on the ordered flight path by controlling pitch, roll, and yaw. The attitude control system operates as an autopilot, damping out fluctuations that would tend to deflect the missile from its ordered flight path. The functions of the flight path control system are to determine the flight path necessary for target interception and to generate the orders to the attitude control system to maintain that path.

It should be clear at this point that the concept of guidance and control involves not only the maintenance of a particular vehicle's path from point *A* to *B* in space, but also the proper behavior of the vehicle while following that path. A weapon that follows a prescribed path half the way to a target and then becomes dynamically unstable or fails structurally due to aerodynamic loading is then incapable of remaining upon the path. Such a vehicle, in order to perform properly, must be "piloted" and capable of responding to control signals.

The operation of a guidance and control system is based on the principle of feedback. Control units make corrective adjustments to the weapon control surfaces when a guidance error is present. The control units will also adjust the control surfaces to stabilize the weapon in roll, pitch, and yaw. Guidance and stabilization corrections are combined, and the result is applied as an error signal to the overall control system.

Sensors

The guidance system in a weapon can be compared to the human pilot of an airplane. As a pilot guides the plane to the landing field, the guidance system "sees" its target. If the target is far away or otherwise obscured, radio or radar beams can be used to locate it and direct the weapon to it. Heat, light, radio waves, television, the earth's magnetic field, inertial navigation, and the Global Positioning System (GPS) have all been found suitable for specific guidance purposes.

When an electromagnetic source is used to guide the weapon, an antenna and a receiver are installed in the weapon to form a sensor that receives guidance information. Weapons that are guided by other than electromagnetic means use other types of sensors, but each must have some means of receiving "position reports."

The type of sensor used is determined by such factors as the environment, maximum operational range, operating conditions, type of information needed, accuracy required, viewing angle, weight and size of the sensor, and type of target and speed.

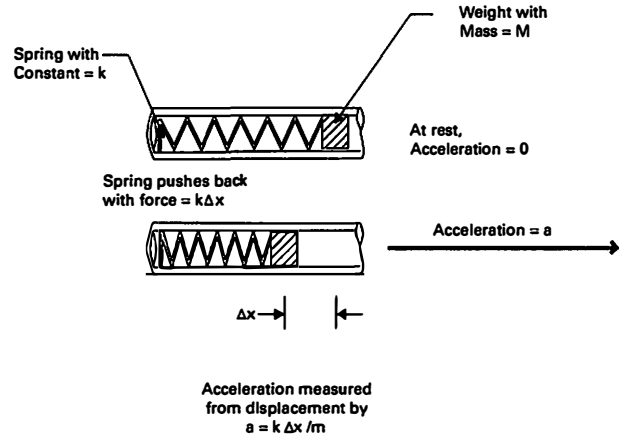


FIGURE 15-1 *Accelerometer*

Accelerometers

An accelerometer, as its name implies, is a device for measuring acceleration. In their basic form, such devices are simple. For example, a pendulum, free to swing on a transverse axis, could be used to measure acceleration along the fore-and-aft axis of the missile. When the missile is given a forward acceleration, the pendulum will tend to lag aft; the actual displacement of the pendulum from its original position will be a function of the magnitude of the accelerating force.

Another simple device is nothing more than a mass, m, on a spring, oriented in the direction of motion, with a scale that measures the compression of the spring; see figure 15-1. The acceleration will cause the spring to compress until the force of the spring matches the mass times the acceleration (recall that $f = ma$). The force of a spring is linear with the change in the length of the spring, Δx, which can be written as $f = k\Delta x$, where k is the spring constant. By measuring the change in length of the spring, the acceleration can be calculated by[i]:

$$a = \frac{k\Delta x}{m} \tag{15-1}$$

If the acceleration along the fore-and-aft axis were constant, the speed of the missile at any instant could be determined simply through multiplying the acceleration by the elapsed time. However, the acceleration may change considerably over a period of time. Under these conditions, integration is necessary to determine the speed.

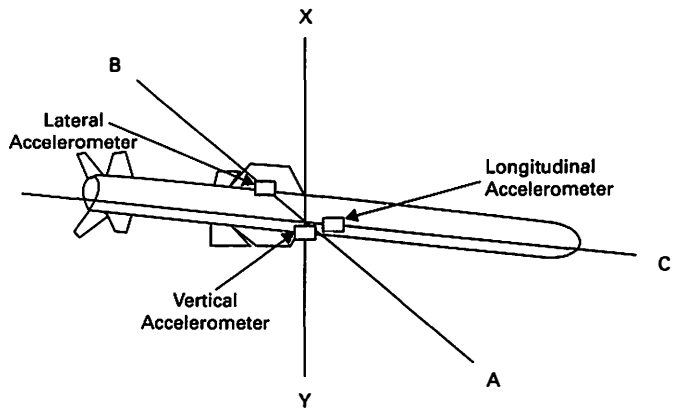


FIGURE 15-2 *Accelerometers in guided missiles*

If the missile speed were constant, the distance covered could be calculated simply through multiplying the speed by time of flight. But because the acceleration varies, the speed also varies. For that reason, a second integration is necessary.

The moving element of the accelerometer is connected to a potentiometer, to a variable inductor core, or to some other device that is capable of producing a voltage proportional to the displacement of the element.

Usually there are three double-integrating accelerometers continuously measuring the distance traveled by the missile in three directions—range, altitude, and bearing, as shown in figure 15-2. Double-integrating accelerometers are devices that are sensitive to acceleration and, by a double-step process, measure distance. These measured distances are compared with the desired distances, which are preset into the missile; if the missile is off-course, correction signals are sent to the control system.

Accelerometers are sensitive to the acceleration of gravity as well as to the acceleration of the missile. For this reason, the accelerometers that measure range and cross-range distances must be mounted in a fixed position with respect to the pull of gravity. This can be done in a moving missile by mounting them on a platform that is stabilized by gyroscopes or star-tracking telescopes.

This stable platform, however, must be moved as the missile passes over the earth to keep the sensitive axis of each accelerometer in a fixed position with respect to the pull of gravity. These factors cause the accuracy of the inertial system to decrease as the time of the missile's flight increases.

To eliminate unwanted oscillations, a damper is included in the accelerometer unit. The damping effort should be just great enough to prevent any oscillations from occurring while still permitting a significant displacement of the mass. When this condition exists, the movement of the mass will be exactly proportional to the acceleration of the vehicle.

Figure 15-3 shows a mass suspended by a spring in a liquid-damped system. If the case experiences acceleration in the direction indicated by the arrow, the spring will offer a restraining force proportional to the downward displacement of the mass, and the viscous fluid will serve to dampen any undesirable oscillations.

Figure 15-4 shows a system that is electrically damped. The mass (M) is free to slide back and forth in relation to the iron core (C). When the vehicle experiences acceleration, the voltage (E), which is proportional to the displacement of the mass, is picked off and amplified.

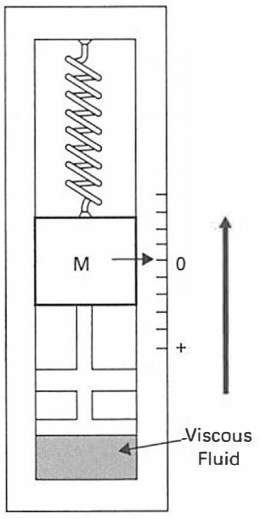


FIGURE 15-3 *Liquid-damped system*

The current (I) (still proportional to the displacement) is sent back to the coil around the core. The resulting magnetic field around the coil creates a force on the mass, which damps the oscillations. In this system the acceleration could be measured by the displacement of the mass (X), by the voltage (E), or by the current (I).

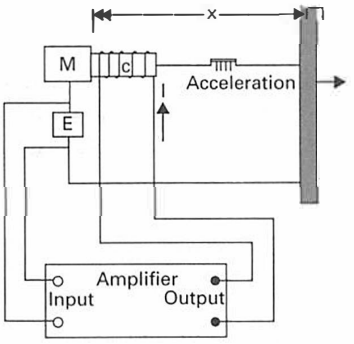


FIGURE 15-4
Electrically damped accelerometer

Phases of Guidance

Missile guidance is generally divided into three phases: boost, midcourse, and terminal; see figure 15-5. These names refer to different parts of the flight path. Not all missiles go through the three phases, but it is a convenient way to divide the weapon's flight.

Boost (Initial) Phase

The boost phase, also called the launching or initial phase, starts when the weapon receives a launch signal

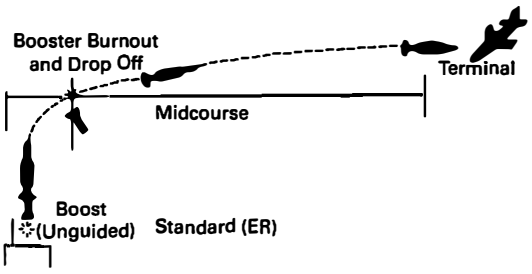


FIGURE 15-5 *Guidance phases of missile flight*

and ends when the weapon is in stable flight toward the target. The specifics are different for each weapon, but generally a weapon is assisted to flight speed by means of a booster component of the propulsion system.

This boost period lasts from the time the weapon leaves the launcher until the booster exhausts its fuel supply. In weapons with separate boosters, the booster drops away from the weapon at burnout, thereby reducing drag on the weapon.

The problems of the boost phase and the methods of solving them vary for different weapons and the different methods of launch; the basic principles, however, are the same. The launcher is aimed in a specific direction on orders from the fire control system. This establishes the line-of-fire (the trajectory or flight path) along which the weapon must travel during the boosted portion of its flight. At the end of the boost period, the weapon must be at the precalculated point in stable flight.

There are several reasons the boost phase is important. If the weapon has homing guidance, it must "look" in a predetermined direction in order to acquire the target. The fire control system calculated this predicted target position prior to the weapon's launch and feeds this information to the weapon on the basis of where the weapon should be at the end of the boost period. Additionally, if the weapon must receive external guidance once launched, the boost phase must end with the weapon in position to receive these signals.

During the boost phase of some missiles, the missile's guidance system and the control surfaces are usually locked into position. The locked control surfaces function in much the same manner as do the tail feathers of a dart or arrow, providing stability and causing the missile to fly in a straight line.

Midcourse Phase

Not all guided weapons have a midcourse phase; but when present, it is often the longest in both time and range. During this part of flight, changes may be needed to bring the missile onto the desired course and to make certain that it stays on that course. In most cases, midcourse guidance is used to put the missile near the target where the final phase of guidance will take over.

Terminal Phase

The terminal or final phase is extremely important. This last phase of guidance must have a high degree of accuracy, as well as a fast response to guidance signals to ensure an intercept. Near the end of its flight, the weapon may be required to maneuver to its maximum capability, sometimes making sharp turns at high speeds to overtake and hit a fast-moving, evasive target.

In some weapons, maneuvers are limited during the early part of the terminal phase. As the weapon gets closer to the target, it becomes more responsive to the positional error signals. In this way, excessive maneuvers are avoided during the first part of the terminal phase, which helps to counter the effect of evasive maneuvering by the target.

Homing Logic

The way in which the weapon closes the distance to the target is called homing logic. There are many approaches to how the weapon should respond in each phase of engagement, each suited to different demands of the target.

A guided weapon is usually under the combined influence of natural and man-made forces during its entire flight. Its path may assume almost any form. Man-made forces include thrust and directional control, as shown in figure 15-6.

The vector sum of all the forces, natural and man-made, acting on a weapon at any instant may be called the total force vector. It is this vector, considered as a function of time in magnitude and direction, that provides velocity vector control and as such depends upon the type of homing logic applied.

Paths along which a guided weapon may travel may be broadly classified as either preset or variable. The plan of a preset path cannot be changed in midflight; the plan of a variable path, however, is altered according to conditions occurring during flight.

Preset Homing Logic

Constant A preset guided weapon path has a plan that has been fixed beforehand. This plan may include several different phases, but once the weapon is launched, the plan cannot be changed. The phases must follow one another as originally outlined. The simplest type of preset guided

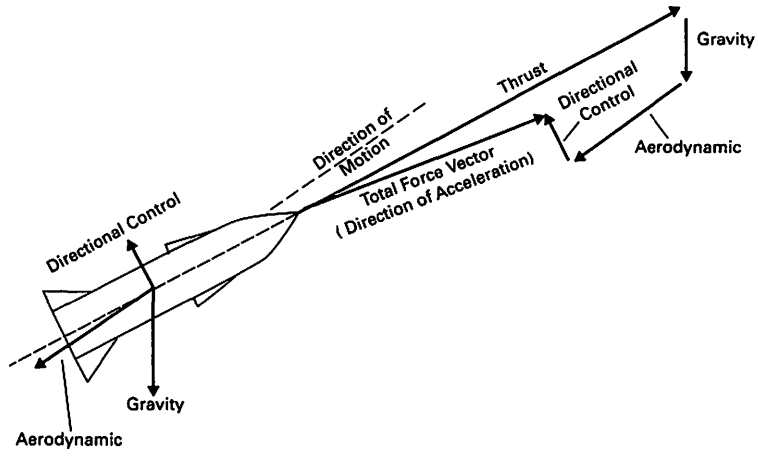


FIGURE 15-6 *Vector sum diagram*

weapon path is the constant preset. Here, the weapon flight has only one phase.

The term *constant preset* may be broadened to include flights that are constant after a brief launching phase that is different in character from the rest of the flight. During the main phase of a constant preset flight, the weapon receives no control except that which has already been built into it.

However, it receives this control throughout the guided phase of flight. Often it is powered all the way. The nature of a constant preset flight path depends on how it is powered and the medium through which the weapon travels.

A torpedo fired from a submarine to intercept a surface target, as in figure 15-7, may describe a straight line—a constant preset guided path.

Programmed A more complex type of preset path is the programmed preset. Here, the weapon flight has several phases. A torpedo, as illustrated in figure 15-8, executes a search pattern. During the first phase, the torpedo, having been launched in some initial direction other than that desired, gradually finds the desired direction by control mechanisms such as gyros and depth settings.

The torpedo then maintains this direction for the remainder of this first phase, at the end of which it is presumed to be in the vicinity of the target and will start the second phase. During this second phase, the torpedo executes a search pattern, possibly a circular or helical path, until it acquires the target and then transitions into the terminal phase.

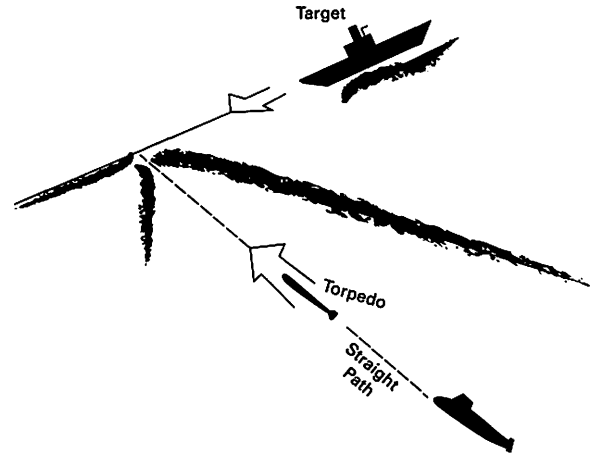


FIGURE 15-7 *Constant preset guided path*

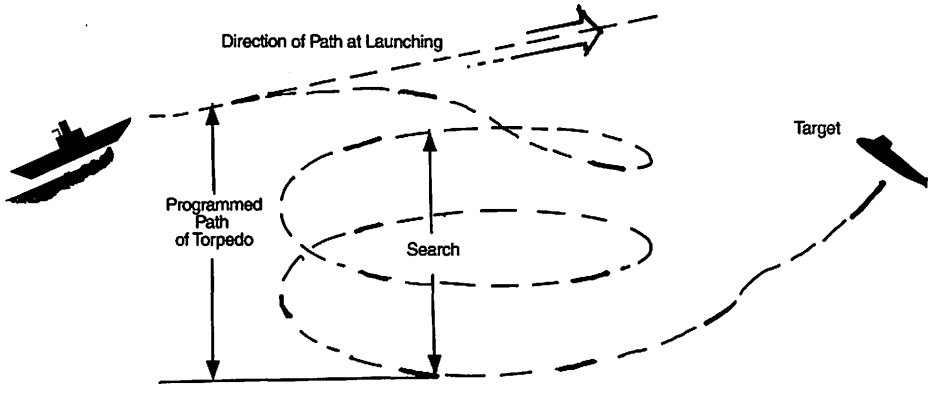


FIGURE 15-8 *Torpedo programmed path*

Variable Homing Logic

Pursuit The simplest path for a guided weapon to follow is to continually fly toward and remain pointed at the target at all times. The weapon is constantly heading along the line-of-sight from the weapon's current position to the target, and its track describes a pursuit path with the rate of turn of the weapon always equal to the rate of turn of the line-of-sight.

Eventually, the flight path will take the weapon either directly ahead or behind the target, usually the latter, which is why this is called pursuit logic. Pure pursuit paths are highly curved near the end of flight, and often the missile may lack sufficient maneuverability to maintain a pure pursuit path in the terminal phase of guidance. When this is the case, the missile can be designed to continue turning at its maximum rate until a point is reached where a pursuit course can be resumed.

The most common application of pursuit logic is against slow-moving targets, or for missiles launched from a point to the rear of the target. This design method is simple to implement but requires a significant speed and maneuverability advantage over the target if the weapon is to close the range quickly.

> *Pursuit:* Lead or deviated pursuit course is defined as a course in which the angle between the velocity vector and the line-of-sight from the weapon to the target is fixed. For purposes of illustration, the lead angle is assumed to be zero, and only pure pursuit is described ($\theta_M = \beta$); see figure 15-9.

Constant bearing At the opposite extreme of a pursuit path is a constant-bearing or collision path. The weapon is aimed at a point ahead of the target, where both the weapon and the target will arrive at the same instant. Recall that this is also the fire control solution for the constant speed and course weapon.

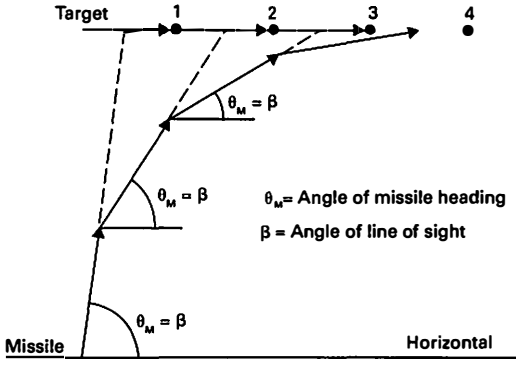


FIGURE 15-9 *Pursuit path*

The line-of-sight to this point does not rotate relative to the weapon. The flight path is as linear as the effects of gravity and aerodynamic forces allow. If the target were to make an evasive turn or if the target's velocity were to change, a new collision course would then need to be computed and the weapon flight path altered accordingly.

The outstanding feature of this type of homing logic is that for a maneuvering constant-speed target, the missile's lateral acceleration never exceeds the target's lateral acceleration. The major drawback, however, is that the control system must have sufficient data-gathering and processing equipment to predict future target position. This method is more suitable for cases where the weapon's speed is comparable to that of the target. As long as the target is not moving directly away, the weapon will close the range quickly.

> *Constant Bearing:* A course in which the line-of-sight from the weapon to the target maintains a constant direction in space. If both weapon and target speeds are constant, a collision course results; see figure 15-10.

$$\frac{\partial \beta}{\partial t} = \beta = 0 \qquad (15\text{-}2)$$

Line-of-sight (LOS) In this path, the weapon is guided so that it travels along the line-of-sight from the launching station to the target. An alternative form of a line-of-sight path is the constant lead angle path. Here, the weapon follows a path that is kept ahead of the line-of-sight by a constant offset.

The major advantages of the line-of-sight path are its flexibility and the minimal complexity of the equipment that must be carried in the weapon, since the major burden of guidance is assumed at the launching station.

> *Line-of-sight:* Defined as a course in which the weapon is guided so that it remains on the line joining the target and the point of control. This method is usually called "beam riding." See figure 15-11.

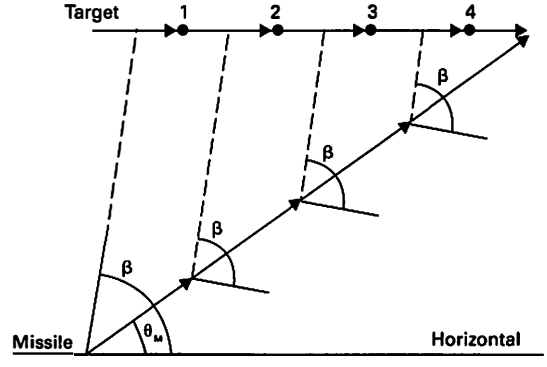


FIGURE 15-10 *Constant-bearing path*

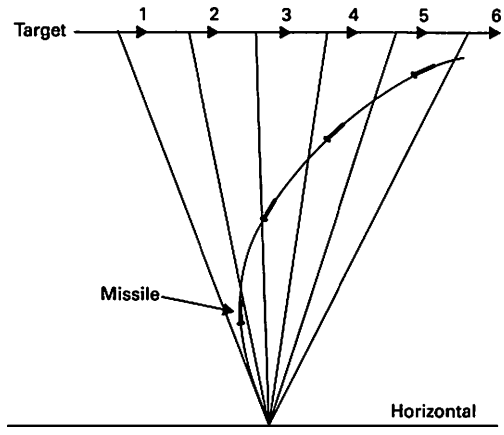


FIGURE 15-11 *Line-of-sight path*

Proportional navigation The more advanced homing weapons employ some form of proportional navigation in which the weapon guidance receiver measures the rate of change of the line-of-sight or bearing drift and passes that information to the guidance computer, which in turn generates steering commands for the autopilot. The weapon rate of turn is some fixed or variable multiple of the rate of change of the LOS. This multiple, called the navigation ratio, can be varied during flight to optimize performance.

A weapon employing this method is said to use a proportional navigation ratio (K), which may be less than 1:1 early in the flight to conserve velocity and increase range. As the flight proceeds, the navigation ratio will increase to 2:1, 4:1, or even greater to ensure that the weapon will be agile enough to counter the target's maneuvers in the terminal phase of flight.

Proportional: A course in which the rate of change of the weapon heading is directly proportional to the rate of rotation of the line-of-sight from weapon to target; see figure 15-12.

$$\frac{\partial \theta_M}{\partial t} = K\frac{\partial \beta}{\partial t} \text{ or } \theta_M = K\beta \qquad (15\text{-}3)$$

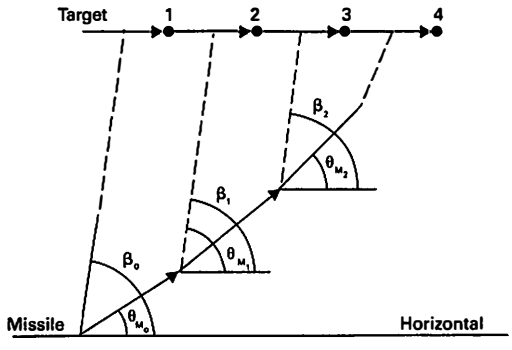


FIGURE 15-12 *Proportional navigation flight path*

Types of Guidance Systems

Guidance systems are what ensure that the weapon is delivered to the target. Guidance systems may be classified into three broad categories with many different functions, some of them overlapping: control guidance, homing guidance, and self-contained guidance.

In the first category are those weapons controlled by radar, radio devices, sonar, or some other means. Homing guidance weapons use the target itself as a source of guidance. In the third category are weapons that rely on electromechanical devices or electromagnetic contact with natural sources such as the stars.

Control Guidance

Control guidance weapons are those that receive commands via a communication path. Control guidance generally depends on the use of radar (radar control), radio (radio control), or wire control links including fiber optics between a control point and the weapon.

This chapter will use radar control guidance as a model for discussion because it is by far the most common application of control guidance methods. The principles discussed may be readily applied to radio (including television) control guidance and wire or fiber-optic guided applications.

Control guidance may be subdivided into two categories. The first category is simply referred to as the command guidance method. The second is the beam-rider method, which is actually a modification of the first, but with the radar being used in a different manner.

Command guidance The term *command* is used to describe a guidance method in which all guidance instructions, or commands, come from sources outside the weapon. The guidance system of the weapon contains a receiver that is capable of receiving instructions from ships, ground stations, or aircraft. The weapon flight path control system then converts these commands to guidance information, which is fed to the attitude control system.

Figure 15-13 is a block diagram of how this method works in actual practice. As soon as the sensor, in this case a radar, is locked on the target, tracking information is fed into a computer. The missile is then launched and is tracked by the same radar, or in some cases a separate radar system.

Target and missile ranges, elevations, and bearings are continuously fed to the computer. This information is analyzed, and a missile intercept flight path is computed on board the weapon control platform. The appropriate guidance signals are then transmitted to the missile receiver.

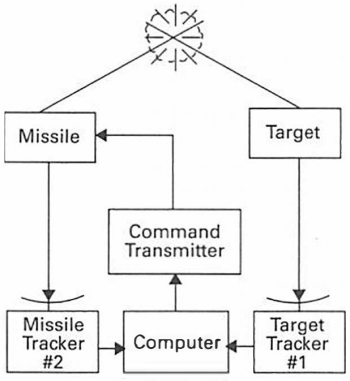


FIGURE 15-13 *Command guidance system*

These signals may be sent by varying the characteristics of the missile tracking radar beam, or by way of a separate radio transmitter. The radar command guidance method can be used in ship, air, or ground weapon delivery systems.

Another application of this type of command guidance is transmitting the commands via a wire or fiber-optical cable. In an antitank role, these systems use an electro-optical sight for tracking the target while the weapon emits a characteristic infrared signature, which is used for tracking the weapon with an IR sensor.

Deviation of the weapon from the line-of-sight to the target is sensed, and guidance commands are generated that are fed to the weapon control system in flight via a direct wire or fiber-optic link. Each weapon contains spools that unwind the transmission wire or fiber as the warhead flies out of the line-of-sight of the target.

Current usage of these systems is in relatively lightweight, portable, short-range battlefield environments against armored targets where these systems' high accuracy is most efficiently employed. An additional use of a wire-guided command system is tying sonar fire control information to a torpedo via a wire link.

The advantage of a wire or fiber link over traditional transmission paths is that the wire is covert and its signals cannot be intercepted or jammed. A disadvantage of a wire/fiber-guided system is that the launch platform must remain relatively stable or even in a fixed position until the weapon impacts the target. While the weapon is en route to the target, depending upon the system, the launcher may also be exposed to enemy counterfire.

Beam-rider method The main difference between the beam-rider method and the radar command guidance method is that the characteristics of the missile-tracking radar beam are not varied in the beam-rider system. The weapon has been designed so that it is able to formulate its own correction signals on the basis of its position with respect to the radar scan axis.

The technique is best understood after reviewing the principles of conical scan tracking in chapter 5. The weapon's flight path control unit is sensitive to any deviation from the scan axis of the guidance radar and is capable of computing the proper flight path correction.

An advantage of this type of system is that it requires only one radar. This radar must, of course, have a conical scan feature in order to provide both target-tracking capability and a weapon flight path correction reference axis. A second advantage is that since the weapon formulates its own directional commands, several weapons may be launched to "ride" the beam simultaneously, without the need for a cumbersome and complicated multiple-weapon command system.

Figure 15-14 illustrates a simple beam-rider guidance system on a typical LOS course. The accuracy of this system decreases with range because the radar beam spreads out, and it is more difficult for the missile to remain in its center. If the target is moving very rapidly, the missile must

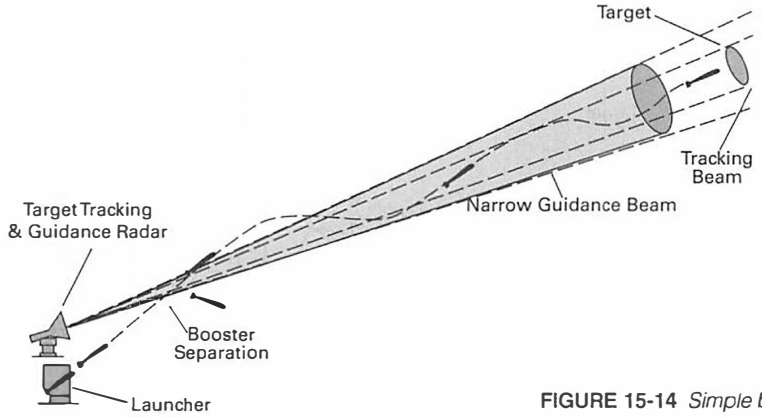


FIGURE 15-14 *Simple beam-rider guidance system*

follow a continuously changing path, which may cause it to undergo excessive transverse accelerations.

Homing Guidance

Homing guidance systems control the flight path by employing a device in the weapon that reacts to some distinguishing feature of the target. Homing devices can be made sensitive to a variety of energy forms including RF, IR, reflected laser, sound, and visible light. In order to home on the target, the missile or torpedo must determine at least the bearing and elevation of the target by one of the means of homing logic mentioned previously.

Some active homing missiles also have the ability to determine the range of the target if necessary. Tracking is performed by a movable seeker antenna or a stationary electronically scanned array.

Homing guidance methods may be divided into three types: active, semiactive, and passive homing, with retransmission homing being a special case; see figure 15-15. These methods may be employed in seekers using any of the energy forms mentioned above, although some methods may be excluded by the nature of the energy form; for example, one would not build a passive laser seeker or an active or semiactive infrared seeker.

Active homing In active homing, the weapon contains both the transmitter and the receiver. Search and acquisition are conducted as with any tracking sensor. The target is tracked employing monostatic geometry in which the returning echo from the target travels the same path as the transmitted energy; see diagram A in figure 15-15. An onboard computer calculates a course to intercept the target and sends steering commands to the weapon's autopilot.

The monostatic geometry allows the most efficient reflection of energy from the target because the small size of the weapon restricts the designer to high frequencies and low power output from the transmitter, which results in short seeker acquisition range.

Semiactive homing In semiactive homing, the target is illuminated by a tracking radar at the launching site or other control point; see diagram B in figure 15-15. The weapon is equipped with a receiver only and by means of reflected radar energy from the target, formulates its own correction signals as in the active method.

However, semiactive homing uses bistatic reflection from the target, meaning that because the illuminator platform and weapon receiver are not co-located, the radar echo which the missile follows is a different path than the energy incident to the target. Additionally, due to its shape and composition, the target may not reflect energy efficiently in the direction of the weapon.

In extreme cases the weapon may lose the target entirely, resulting in a missed intercept. This disadvantage is compensated for by semiactive homing's ability to use greater power and more diverse frequency ranges in an illumination device on a ship, aircraft, or ground station and also less signal loss due to path length.

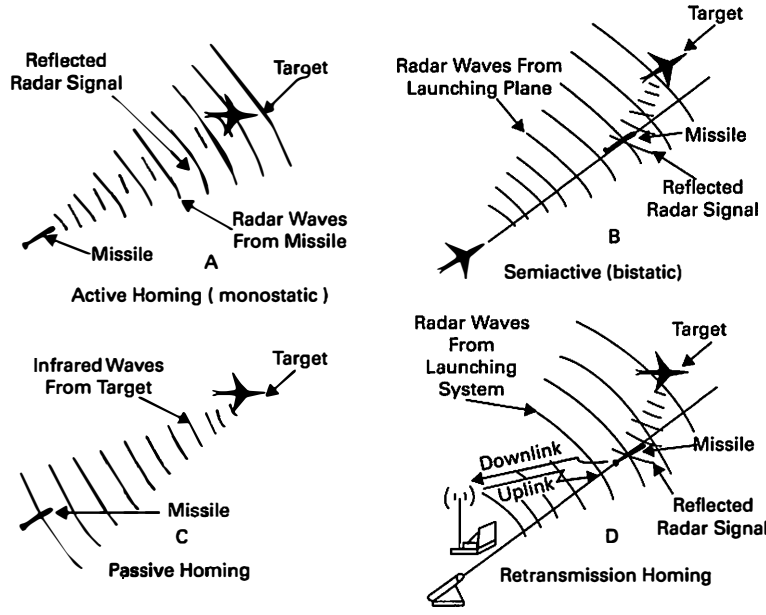


FIGURE 15-15 *Homing guidance methods*

Passive homing Passive homing depends only on the target for a source of tracking energy; see diagram C in figure 15-15. This energy can be the noise radiated by a ship or submarine in the case of a passive homing torpedo; radio frequency radiation from the target's own sensors in the case of an antiradiation (ARM) weapon; heat sources such as ship, aircraft, or vehicle exhausts; contrast with the temperature or visible light environment; or even the radiation all objects emit in the IR region.

As in the other homing methods, the missile generates its own correction signals on the basis of energy received from the target rather than from a control point. The advantages of passive homing are that counterdetection is reduced and that a wide range of energy forms and frequencies are available. Its disadvantages are its susceptibility to decoy or deception and its dependence on a certain amount of cooperation from the enemy.

Retransmission homing or track via missile (TVM) Retransmission homing is a blend of the characteristics of both command and semiactive homing guidance; see diagram D in figure 15-15. In command guidance, missile-steering commands are computed at the launch point using target position and missile position data derived from launch point sensors.

In retransmission homing, the missile contains a semiactive seeker that determines the bearing and elevation angle from the missile to the target, which is then coded and transmitted to the launch point via data link (downlink). The fire control system at the launch point can use its own target tracking data, that of the missile, and missile position data to compute steering commands, which are then transmitted to the missile via an uplink.

This technique is used in the more advanced air defense missile systems, including the U.S. Navy Standard missile and the U.S. Army Patriot system. Specific retransmission or TVM systems may vary somewhat from this ideal; however, they all in some way will use target angle data from the missile to compute steering commands at the launch point that are then transmitted to the missile.

Accuracy Homing, when used against moving targets, is the most accurate of all guidance systems because it uses the target as its source. There are several ways in which the homing device may control the path of the weapon against a moving target. Of these, the more generally used are pursuit paths and lead flight paths.

Because monopulse methods in weapon seekers are advantageous and are becoming the method of choice in current weapons, it is necessary to address the two basic types.

Amplitude comparison monopulse This method, described in chapter 5, requires a gimbaled seeker antenna covered by a radome at the nose of the weapon. Because of aerodynamic requirements, the radome shape is normally not optimal for radar performance. Very precise orders to the antenna are required to achieve target acquisition due to the single antenna's limited field of view.

In these systems, the size of the antenna directly determines the limits of the frequency range of the seeker. Its primary advantage is its consistent performance throughout the speed and maneuverability range of potential targets.

Interferometer (phase comparison monopulse) The interferometer eliminates the requirement for a movable antenna, because fixed antennas are instead mounted at the edge of the airframe or on the wingtips. The result is reduced complexity and a wider field of view.

As depicted in figure 15-16, two antennas separated by a known distance are installed for each mobility axis of the weapon. In the diagram, antennas A and B, separated by the distance d, receive energy emitted (passive homing) or reflected (semiactive homing) from the target.

Because the distance to the target is relatively large, it is assumed that the RF energy arrives as a series of planar waves with wavelength λ. In accordance with the discussion of electronic scanning in chapter 4, it is evident that for the geometry pictured, the phase sensed by antenna B will lag that sensed by antenna A by some phase angle θ that is proportional to $d \sin \beta$. Therefore:

$$\theta = \frac{2\pi d}{\lambda} \sin\beta \qquad (15\text{-}4)$$

If λ is known and the phase angle θ can be determined, then the look angle, β, can be calculated.

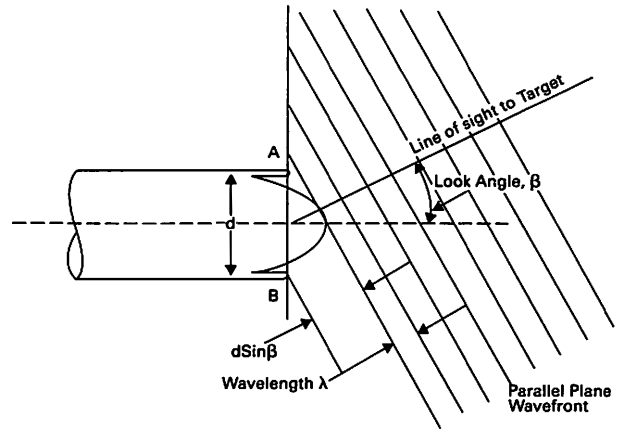


FIGURE 15-16 *Body-fixed RF interferometer*

The interferometer provides the advantages of a wide field of view, flexibility in airframe design, unobstructed use of weapon interior space, and the ability to cover broad frequency bands without constraints imposed by limited antenna size. The separation between the antennas governs the performance of the system, with missile body diameter or fin spread separation as the usual arrangement.

The disadvantage of the interferometer is the angular ambiguity that may exist for wavelengths smaller than the separation between the antennas at a specific angle of incidence. If the distance between the antennas at an angle of incidence β is $d \sin \beta$, and λ is less than $d \sin \beta$, then it is not possible to determine whether is the phase angle sensed is as measured or $\theta + n2\pi$ radians, where n is any integer. However, this is a minor problem in most homing systems because the absolute look angle is not as important as the rate of change of that angle.

The interferometer has an advantage in resolving multiple targets at twice the range of a typical amplitude comparison monopulse seeker in the same size weapon. This gives the missile twice the time to respond to the changeover from tracking the centroid of the group to tracking one specific target, thus increasing the hit probability.

Composite systems No one system is best suited for all phases of guidance. It is logical, then, to combine a system that has good midcourse guidance characteristics with a system that has excellent terminal guidance characteristics, in order to increase the probability of hit. Combined systems are known as composite guidance systems or combination systems.

Many weapons rely on combinations of various types of guidance. For example, one type of weapon may use command guidance until it is within a certain range of a target, at which time the command guidance may become a backup mode and a type of homing guidance commences. The homing guidance will then be used until impact with the target or detonation of a proximity-fuzed warhead.

This type of combination achieves many advantages of both systems. It attains long-range capabilities by maintaining the tracking sensors on the delivery vehicle, be it a ship, aircraft, or ground station, and by transmitting the data to the weapon. By having the weapon compute its own attitude adjustments, the fire control problem is thereby simplified.

Self-Contained Guidance

Self-contained guidance falls in the second category of guidance system types. In this type of guidance, all the control equipment is entirely within the missile. Examples of this type are preset, navigational, inertial, celestial, terrestrial, and GPS. These systems are most commonly applicable to weapons designed to hit fixed targets and also where the designer desires to minimize the effects of electronic countermeasures since these guidance systems (with the exception of GPS) neither transmit nor receive signals that can be jammed.

Preset guidance The term *preset* completely describes one guidance method. When preset guidance is used, all of the control equipment is inside the weapon. This means that before the weapon is launched, all information relative to target location as well as to the trajectory the weapon must follow is calculated.

After this is completed, the weapon guidance system is set to follow the course to the target, hold the weapon at the desired altitude, measure its speed, and, at the correct time, cause the weapon to start the terminal phase of its flight and pitch toward the target.

A major advantage of preset guidance is that it is relatively simple compared to other types of guidance because it does not require tracking or visibility.

An early example of a preset guidance system was the German V-2, where range and bearing of the target were predetermined and set into the control mechanism. The earliest Polaris missile was also designed to use preset guidance during the first part of its flight, but this was soon modified to permit greater launch flexibility.

The preset method of guidance is useful only against stationary targets of large size such as landmasses or cities. Since the guidance information is completely determined prior to launch, this method is not suitable for use against ships, aircraft, enemy missiles, or moving land targets.

Navigational guidance systems When targets are located at great distances from the launching site, some form of navigational guidance must be used. Accuracy at long distances is achieved only after exact and comprehensive calculations of the flight path have been made.

The mathematical equation for a navigation problem of this type may contain factors designed to control the movement of the missile about the three axes of pitch, roll, and yaw. In addition, the equation may contain factors that take into account acceleration due to outside forces—for example, tailwinds and the inertia of the missile itself.

Four navigational systems that are used for long-range missile guidance are inertial, celestial, terrestrial, and the Global Positioning System (GPS).

Inertial guidance The simplest principle for guidance is the law of inertia. In aiming a basketball at a basket, an attempt is made to give the ball a trajectory that will

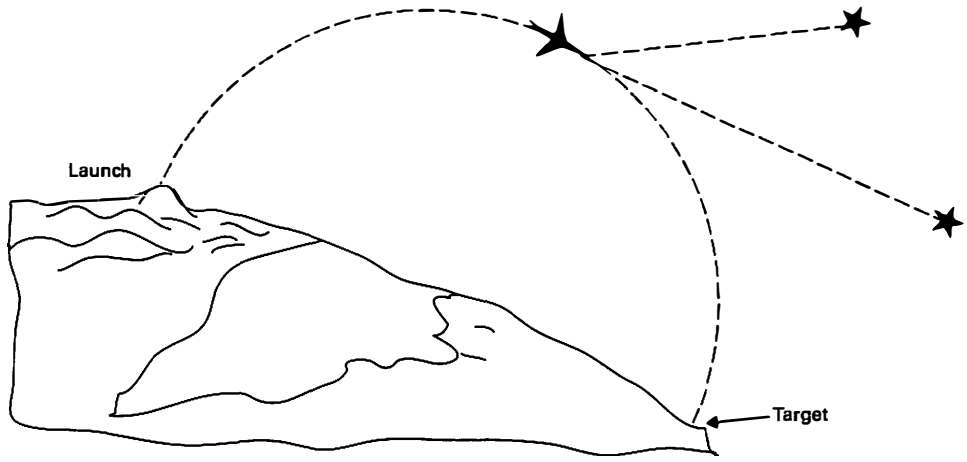


FIGURE 15-17 *Celestial guidance*

terminate in the basket. However, once the ball is released, the shooter has no further control over it. If the shooter has aimed incorrectly, or if another person touches the ball, it will miss the basket.

However, it is possible for the ball to be incorrectly aimed and then for another person to touch it and change its course so that it will hit the basket. In this case, the second player has provided a form of guidance. The inertial guidance system likewise supplies the intermediate push to get the missile back on the proper trajectory.

The inertial guidance method is used for the same purpose as the preset method and is actually a refinement of that method. The inertially guided weapon also receives programmed information prior to launch. Although there is no electromagnetic contact between the launching site and the weapon after launch, the weapon is able to make corrections to its flight path with amazing precision by controlling its flight path with accelerometers that are mounted on a gyro-stabilized platform.

All in-flight accelerations are continuously measured by way of this arrangement, and the weapon attitude control generates corresponding correction signals to maintain the proper trajectory. The use of inertial guidance takes much of the guesswork out of long-range weapon delivery. The unpredictable outside forces affecting the weapon are continuously sensed by the accelerometers. The generated solution enables the weapon to continuously correct its flight path. Thus the inertial method has proved to be a very reliable long-range guidance method.

Celestial reference A celestial navigation guidance system is a system designed for a predetermined path in which the weapon course is adjusted continuously by reference to fixed stars. The system is based on the known apparent positions of stars or other celestial bodies with respect to a point on the surface of the earth at a given time.

Navigation by fixed stars and the sun is highly desirable for long-range weapons since its accuracy is not dependent on range. Figure 15-17 sketches the application of the system as it might be used for a guided missile.

The weapon must be provided with a horizontal or a vertical reference to the earth, automatic star-tracking telescopes to determine star elevation angles with respect to the reference, a time base, and navigational star tables that have been mechanically or electrically recorded.

A computer in the weapon continuously compares star observations with the time base and the navigational tables to determine the weapon's present position. From this, proper signals are computed to steer the weapon correctly toward the target.

The weapon must carry all this complicated equipment and must fly above the clouds to ensure star visibility. ICBM and SLBM (Submarine Launched Ballistic Missile) systems currently use celestial guidance.

Terrestrial guidance methods Prior to the microminiaturization of computer circuits, the various proposed methods of terrestrial guidance had significant limitations. These suggested early systems included an inertial reference system, a television camera to provide an image of the earth's surface, and a filmstrip of the intended flight path. This guidance system would compare the television picture with the projected filmstrip image and determine the missile's position by matching the various shadings in the two images.

This method proved too slow in providing position data, even for a subsonic missile. Its other distinct disadvantage was that it required extensive low-level aerial

photography of each potential missile flight path. The danger to flight crews and the loss of the element of surprise involved in extensive prestrike photo reconnaissance made such a system impractical.

With the availability of compact mass memory and vastly increased computational capability, both of which are ideal for weapon space and weight limitations, terrestrial guidance methods became practical. The advent of small radar altimeters of high precision provided an alternative to photographic methods, with the added advantage that weather and lighting conditions were relatively inconsequential. The radar altimeter provides a coarse means of detecting surface features by their height, which can then be compared with stored data concerning expected land contours along the missile flight path.

The missile guidance system contains expected land elevation values to the left and right of the weapon's intended ground track. The guidance system will determine when the weapon is located at a position where the stored data most closely match the observed altitudes. Once the direction of turn and distance required to correct the error have been determined, the weapon turns to resume the intended track.

This method is called terrain contour matching, or TERCOM. Even the most capable TERCOM system has insufficient memory to perform contour matching throughout a flight path of several hundred miles, however. Therefore, the weapon is provided with a series of small areas known as TERCOM maps en route to the target. The number of TERCOM maps and their separation are determined by the quality of information available on the area and the accuracy of the missile's inertial navigation system.

Sufficient data to support TERCOM are available from various sources such that aerial reconnaissance of most target areas is not required prior to the engagement. TERCOM has sufficient accuracy to find, for example, a large military base within a region; however, it cannot provide the accuracy for a missile to be able to hit a specific section of that base, such as a group of hangars at an airfield. For this reason, a missile using only some variation of TERCOM would require a nuclear warhead.

Delivery of a conventional high-explosive warhead requires precision that can be provided only by some form of optical device in the terminal stage of flight. A cruise missile flies at altitudes and ranges that would prevent the transmission of images back to the launch point. Advances in digitized imagery, however, permit computer storage of gray-shaded scenes in the vicinity of the target.

The digitized scene can then be compared to data from a television camera in the missile, and values of gray shading can be matched to determine the missile's actual position relative to its desired position, as pictured in figure 15-18. The missile can correct its flight path to that desired position and even finally pick out its target.

This method, called digital scene matching area correlator, or DSMAC, is sufficiently accurate to permit the use of a conventional high-explosive warhead. The DSMAC technique is used only for the last few miles to

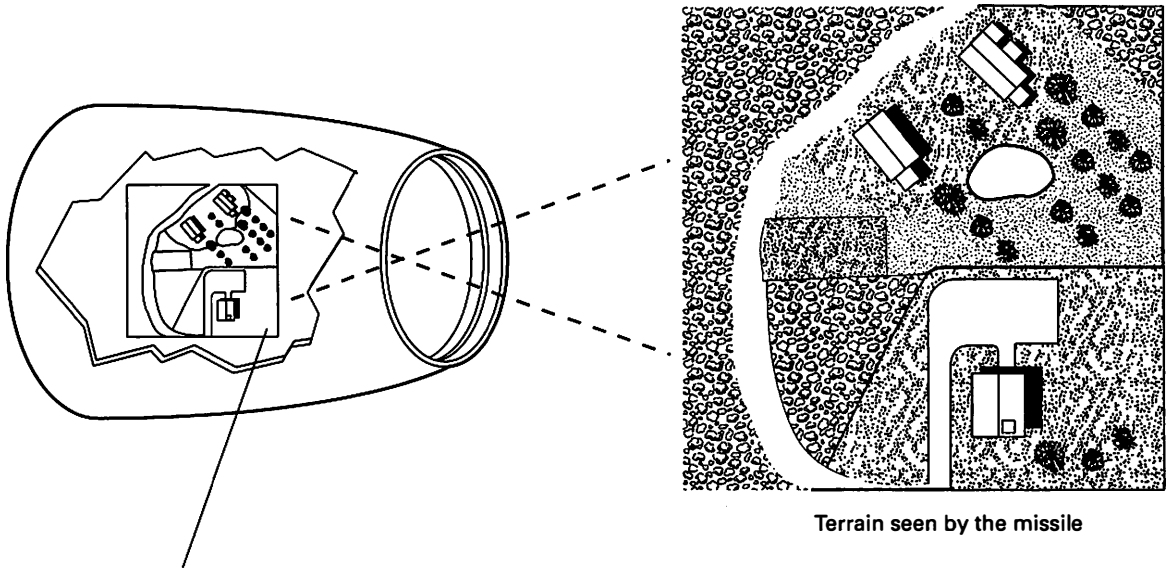


FIGURE 15-18 *Terrestrial guidance*

the target, with the TERCOM method being used for the majority of the flight path.

Both methods are limited by the accuracy of the information used to create the digital TERCOM maps and DSMAC scenes that are loaded into the missile's memory. Building and formatting these data files for cruise missiles thus requires considerable support facilities and talented personnel.

Global positioning system The Global Positioning System (GPS) is a satellite-based navigation system that uses a digital signal on a carrier frequency of 1.5 GHz from each of up to 24 satellites. With sufficient navigational data, the receiver can determine its exact range from each satellite as well as each satellite's geographic position (GP). The GP is the location on the earth that is directly below the satellite. This establishes a line-of-position (LOP) on the earth, as illustrated in figure 15-19.

A second LOP from another satellite will narrow the location of the receiver to two possible locations, as shown in figure 15-20. Finally, a third LOP will resolve the location of the receiver to a unique position on the earth.

The range from the satellite is determined by the time the signal is received. The satellite signal includes the time at which it was sent, and by comparing that to the receiver clock, the time delay and hence the range can be determined.

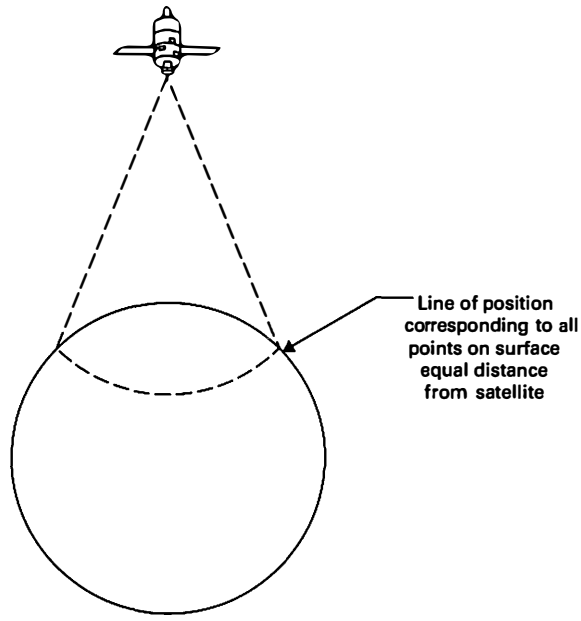


FIGURE 15-19 *Line-of-position on the earth's surface from a single satellite*

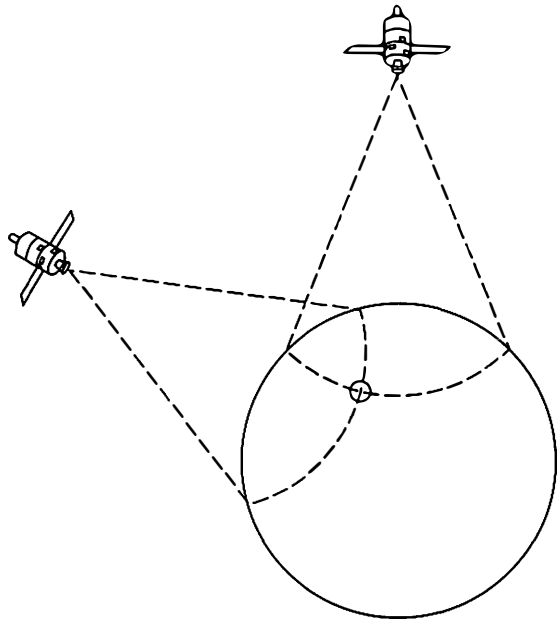


FIGURE 15-20 *Fix from two lines-of-position on the surface of earth (two-dimensional)*

The satellites contain precise atomic clocks that are updated from a master station. The GPS receiver, however, does not contain a precise atomic clock. To determine the correct time, the receiver uses a signal from a fourth satellite. This signal is used to resolve any error in the receiver's clock. As a result, the receiver obtains an accurate three-dimensional navigational fix, as well as the precise time.

In order for the GPS to work, there must be at least four satellites visible to the receiver. In fact, there are a total of 21 operational satellites, with three spares. These satellites circle the globe every 12 hours in polar orbits at an altitude of about 20,000 kilometers.

Assume the 21 satellites are evenly spaced over the 4π steradians of the earth's sphere. (A steradian is the unit of solid angle found by taking the surface area of the sphere divided by the radius squared.) This gives about $4\pi/21 = 0.6$ steradian per satellite. Also, assume the receiver can see about one-quarter of the sky at any one time, about π steradians; that would then involve $\pi/0.6 \approx 5$ satellites. Therefore the constellation of 21 satellites is adequate for providing continuous, worldwide coverage.

The basic system is divided into three segments: space-based, control, and user. The space-based segment consists of the 24 satellites in polar orbits, as shown in figure 15-21. There are usually several inactive spares in orbit at any one time to replace satellites as they reach the end of their useful lives.

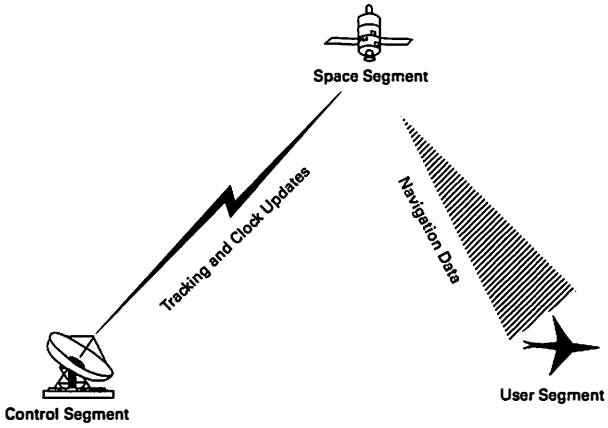


FIGURE 15-21 *Segments of GPS*

At 20,000 kilometers above the earth, the geographic position of the satellites travels along a meridian. The orbits of the various satellites are distributed by increasing their orbital plane in increments of 55°.

The control segment consists of the master control station (MCS) in Colorado Springs, Colorado, and three unmanned stations in Hawaii, Kwajalein, and Diego Garcia, which track the satellites and relay information back to the MCS. There are three uplink antennas, one each at Diego Garcia, Kwajelein, and Ascension island. The control segment tracks the satellites and updates their orbital parameters and internal clocks.

The user segment consists of the GPS receivers. They receive the satellite transmissions and perform calculations to determine position, altitude, and time outputs.

Receiver antennas are designed to receive the circularly polarized satellite signal at either 1,575.42 (L1) or 1,227.60 (L2) MHz. The difference between these signals will be discussed below. The signal is basically composed of a 50 bps bipolar phase-shift-keyed (BPSK) digital signal containing navigation data. This in turn is carried on a 1,023 MHz spread spectrum signal, which serves to encode the identity of the satellite and to widen the signal bandwidth, which reduces the threat of interference from narrowband sources.

The GPS is accurate to within about 53 feet in position, 0.3 fps (feet per second) in velocity, and 100 nanoseconds in time. However, to deny the precise navigational location to possible hostile forces, a random error of anywhere between 130 and 330 feet can be intentionally inserted. The unadulterated signal is known as the PPS (precise positioning system) and when activated is available only to government-approved users who have a special decryption system.

To compensate for the injected errors, a system has been developed outside of the government program to measure and correct for the error. This system is known as differential GPS, or DGPS. Its principle of operation is that a fixed site, with a well-known position, obtains a GPS fix, measures the error, and broadcasts the correction to other DGPS users.

More recently, methods such as the wide area augmentation system (WAAS) have been developed that use geosynchronous satellites to transmit corrections to users. Such corrections enhance accuracy in a manner similar to DGPS, but the area covered is significantly expanded (essentially, the footprint of the satellite).

Despite the greatly improved accuracy of GPS, when attempting to guide a weapon to a target, an error due to how coordinate systems are generated enters the situation. This error has always existed, but up to this point, it was not an issue due to weapon accuracy. With the advent of GPS, the target location error (TLE) must now be accounted for.

Target coordinates are usually based on stereoscopic imagery that is translated into coordinates such as latitude, longitude, and elevation. There will always be some small differences between the location the GPS identifies and the target's actual location; this is the target location error.

Summary

The three phases of guidance are boost, midcourse, and terminal. The distinction between them is primarily based upon a breakdown of the flight path rather than on any changeover points in guidance methods. The terminal phase, however, is the most critical and demands peak performance from the guidance system.

Guidance systems are divided into two broad categories: those that use man-made electromagnetic devices and those that use some other means. The various subcategories of guidance systems are shown in table 15-1.

Guided missile paths may be classified as either preset or variable. Preset guided paths have planned flight routines that cannot be changed in midflight on the basis of updated data. A preset plan may be for a one-phase flight (constant preset) or a flight of several phases (programmed preset).

Variable guided flight paths have plans that can be changed in midflight; thus they make possible the successful intercept of a target that makes evasive maneuvers. Prediction of target position is continuously reappraised, and the weapon's course is recomputed in the light of new targeting data.

TABLE 15-1 *Guidance system categories*

Guidance Systems that Use Man-Made Electromagnetic Devices		**Self-Contained Guidance Systems**	
Control	*Homing*	*Preset*	*Navigational*
Command	Active		Inertial
Beam-rider	Semiactive		Celestial
Modified beam-rider	Passive		Terrestrial
			GPS
Composite Systems			
TVM (Command/Semiactive)			

Variable guided flight paths include pursuit, constant bearing, line-of-sight, and proportional navigation.

A missile's interception of a moving target depends on the prediction of future target position and requires certain assumptions. When using bullets, ballistic missiles, or preset guided missiles, it is assumed that the target motion measured while tracking will remain unchanged during missile flight. When using variable guided weapons, it is assumed that the target motion, measured at any instant by almost continuous tracking, will remain unchanged over a short time interval.

[i] Hall, 249

16 Fuzing

Introduction

A fuze is a weapon subsystem that activates the warhead mechanism in the vicinity of the target and also maintains the warhead in a safe condition prior to use.

The fuze is essentially a binary state mechanism—that is, it is either safe or armed. In the context of weapon system hardware, a fuze and a warhead are unique in that they are expected to remain functionally dormant until a bona fide target is encountered and then to function as intended in a fraction of a millisecond.

Guidance systems may recover from transient malfunctions; target-tracking radars may experience numerous false alarms without significantly compromising their utility; and missile airframes may flex and recover. But the fuze-warhead process is singular and irreversible.

The quality required of fuze designs is usually specified by two values: functional reliability, which ranges typically from 0.95 to 0.99 for complex missile fuzes, and a reliability of 0.999 for projectile and bomb contact fuzes; and safety reliability, for which a failure rate not greater than 1 in 10^6 must be proved prior to service usage.

Function of the Fuze System

The weapon fuze system has five basic functions that it can perform:

1. Keeping the weapon safe
2. Arming the weapon
3. Recognizing or detecting the target
4. Initiating the detonation of the warhead
5. Determining the direction of detonation (special fuzes only)

A typical fuze system that incorporates these five basic functions is depicted in figure 16-1. These functions will be described in greater detail in the following sections.

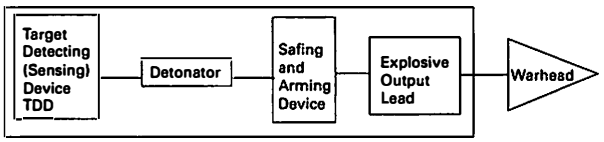


FIGURE 16-1 *Basic fuze system*

Safing and Arming

When prehistoric man first dropped his club on his toe, he realized that his own weapon could be dangerous to himself as well as to his enemies. With today's highly destructive weapons, there must be a high degree of assurance that the weapon will not detonate until it has reached the target that it is intended to destroy. This assurance is provided by the safing and arming (S&A) device.

The S&A device is a fuze component that isolates the detonator from the warhead booster charge until the weapon has been launched in the intended mode and is a safe distance from the launch vehicle. Then, at a programmed point in the trajectory, the S&A device removes a physical barrier from the explosive train, allowing the detonator thereafter to initiate the warhead.

The design of an S&A device must meet several specific common requirements:

1. The device should not utilize energy stored prior to launch as an arming agent.
2. For maximum safety the device should be designed so that the energy required for arming is derived from environmental parameters that exist during normal flight.
3. The device should not be initiated by parameters that exist outside the actual warhead.
4. The warhead must be protected against malfunctioning by a primary safety mechanism.

5. The process of removing the safety features built into a warhead must occur after a unique sequence of events.
6. A secondary safety system should be included if the need arises to combat or neutralize the effects of countermeasures.
7. Redundancy in design is needed to ensure a high degree of proper functionality.

Safing and arming devices use various methods to perform their vital function. Some S&A devices function by measuring elapsed time from launch; others ascertain distance traveled from the launch point by sensing and performing a double integration of the acceleration experienced by the weapon. Still other devices, such as those found in ballistic missiles, use deceleration during the reentry phase. Depth charges and mines may use the change in ambient pressure to initiate arming.

Additionally, some devices sense and integrate airspeed or projectile rotation (centrifugal force), and still others may sense several stimuli characteristic of the weapon trajectory such as rocket motor gas pressure and acceleration.

To maximize the safety reliability of a fuze, the S&A device must ensure that the forces it senses will be unique to the weapon when deployed and cannot be intentionally or accidentally duplicated during ground handling or prelaunch operations.

Most missiles incorporate an S&A device that utilizes a form of acceleration sensing. This device senses the acceleration from the rocket motor. The rocket motor booster may induce only a 10 g force, whereas the weapon dropped from a sling may easily experience 500 g shock forces.

Therefore, another feature must be used to prevent arming should the weapon be dropped. This necessary attribute is that acceleration-sensing systems must experience acceleration for a protracted period of time, such as for at least 2 seconds. Such a system is illustrated in figure 16-2.

Figure 16-2 shows a linear motion g-weight S&A device that drives the "out-of-line" explosive lead into alignment at a rate controlled (typically) by a Geneva escapement, the rate of which is not very sensitive to the acceleration magnitude.

The "intent to launch" signal operates the armature of the solenoid, retracting the launch latch as the weapon experiences forward acceleration. The g-weight tends to move aft against the spring, but its rate of movement is retarded by the Geneva escapement, which functions like the escapement of an alarm clock.

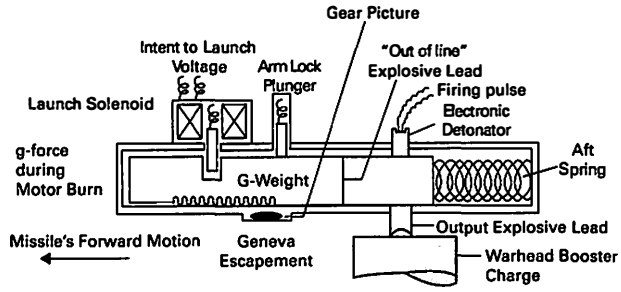


FIGURE 16-2 *Example demonstrating the principle of an acceleration-integration S&A device*

After a period of time, the detonator, the "out-of-line" explosive lead, and the output explosive lead align and form a continuous explosive chain from the detonator to the warhead booster charge. When the g-weight reaches that position, a second pin is pushed by a spring into a detent in the g-weight and locks it in the armed position.

If the rocket motor is faulty or if sufficient acceleration is not sustained for the expected time, then the aft spring will return the g-weight rapidly to its original position (the Geneva escapement has, in effect, a ratchet that permits rapid reverse movement), and the launch latch will be inserted into the g-weight again, locking it into the safe position. This ensures that the warhead will not detonate if it unexpectedly falls short due to motor failure.

The linear motion g-weight acceleration integrator as described is very rugged but has the functional disadvantage that during the lateral acceleration of the missile, which may exceed 30 g in magnitude, the frictional force between the g-weight and the tube increases greatly. This retards the axial motion of the g-weight and extends the arming distance.

In a more efficient design, linear motion is converted into rotary motion of the g-weight. A metal disk mounted eccentrically on a transverse shaft rotates as the missile accelerates, aligning the explosive train.

To obtain the desired safety reliability in the linear example, there are several redundant components: the launch latch, the Geneva escapement, and the g-weight with its spring. If any component were to fail—for example, should the launch latch shear when the weapon is dropped—another component would still prevent inadvertent arming. This is the type of redundancy that is found in all S&A devices.

S&A devices are made of extremely high-quality components and are designed to have a safety failure probability of not greater than 1 in 10^6. This degree of safety is the usual level required of Navy missile S&A devices.

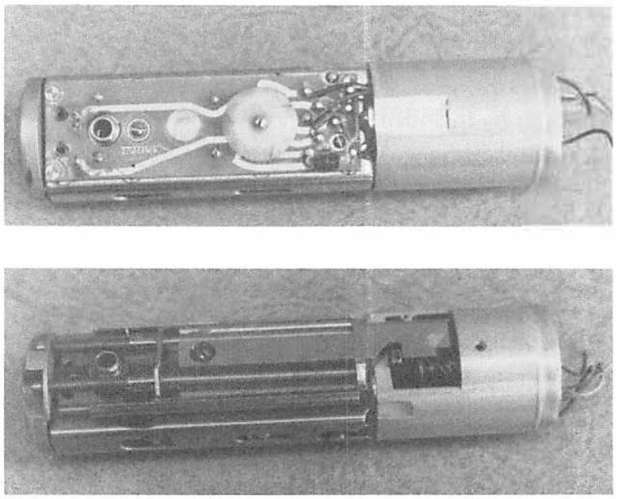


FIGURE 16-3 *Safing and arming device*

The discussion so far has emphasized keeping the weapon safe before it reaches the target. It may also be necessary to include an additional S&A device in the firing path in order to safe the weapon if the intended target is missed. An example would be an antiaircraft projectile or SAM that could fall on friendly territory if an airborne target were missed.

To preclude this possibility, a surface-to-air missile could have an additional timer S&A-type device that would normally arm. This would allow firing path continuity until the expected elapsed time from launch to target intercept is exceeded, at which time the device will safe the missile.

An alternative to using a timer S&A device would be to place an additional timer firing device in the fuze system in addition to the normal target-sensing device. If the target is missed, then the alternate timer device would send the fire signal for detonation. The timer setting would be based on a time greater than normal intercept, but would detonate while the weapon was still safely airborne. This alternative design method of increasing the safety reliability is found in most antiaircraft projectiles.

Although the fuze with its S&A device is the primary safety component of the weapon, it must be realized that the warhead is sensitive to influences other than fuze initiation. The chemical high explosives used in both conventional and nuclear warheads can be detonated if subjected to enough heat and energy from external sources. Combat or accidental detonations or fire can provide this required energy.

Target Detection and Recognition

Target detection and recognition must occur either when the target is close enough to be contained within the damage volume of the warhead or when the proximity fuze can predict when this condition will occur. This function can be accomplished by several types of mechanical or electrical sensor devices. The specific device used classifies the type of fuzing system. Once the target has been detected or "sensed," the detection and recognition device either sends, or predicts when to send, a fire signal to the detonator. There are four basic categories of these devices.

Impact or contact This sensing device detects the weapon's first impact with an object. Its output initiates the detonator. Typical sensor mechanisms include displacing an inertial mass, stressing a piezoelectric crystal, short-circuiting a coaxial transmission line, and interrupting an electrical circuit. Point detonating fuzes are of this type.

Delay mechanisms can also be incorporated that provide a short detonation delay, usually in microseconds, after weapon contact and allow for weapon penetration of the target to maximize damage effect.

Ambient Although not able to detect the physical presence of the target, this type of device can sense the unique environment in which the target can be found, such as water depth. This type of sensing device would normally be found on depth-bomb and depth-charge weapons.

Additionally, air pressure can be used to determine burst height for large blast weapons such as fuel-air-explosive and nuclear bombs.

Timer/command After a predetermined time has elapsed, a timing device will send a fire signal to the detonator. The "sensing" or "detecting" of the target is predetermined by the fire control system in that a preset elapsed time is based on calculations of when the target is expected to be within the damage volume. Gun projectiles often use this type of fuzing. Bombs can have time fuzes that can be set to go off from minutes to hours after striking the ground. See figure 16-4.



FIGURE 16-4 *An aircraft bomb time fuze*

Proximity This device does not require target contact but can sense the physical presence of the target at some distance away. It sends a fire signal to the detonator when the target is within the predicted damage volume of the warhead. Proximity fuzes are called target detecting devices, or TDDs and will be further investigated later in the chapter.

Warhead Initiation

A warhead usually contains a powerful but relatively insensitive high explosive that can be initiated only by the heat and energy from a primary explosive. The primary explosive is a component of the fuze subsystem and is normally loaded in the detonator. If the detonator is designed properly, it can be activated only by a unique fire signal received from the target detection and recognition system. A detonator can be designed to activate when it receives either electrical energy (high voltage) or mechanical energy (shock or stab) from the target sensor.

Fuze System Classification

Fuze designs vary widely according to weapon characteristics and mission. Of the several naming conventions employed in the Navy, a convenient system is to classify fuzes by their manner of operation:

1. Proximity fuzes
   a. Active
   b. Semiactive
   c. Passive
2. Timer/command fuzes
3. Impact or contact fuzes
4. Ambient

Proximity Fuzes

By far the most complex class of fuze is the proximity type. Proximity fuzing had its origins in the United States and England early in World War II. Operations research analysts, pondering how to increase the effectiveness of their antiaircraft artillery, calculated that a proximity-sensing fuze carried on a projectile could make an enemy bomber effectively ten times as large as its physical size by detonating the projectile in the vicinity of the target rather than at a fixed flight time of the projectile.

Thus, of those projectiles that would intercept the target at a range where damage could be inflicted, ten times as many would be effective in damaging or destroying the target in comparison to older time-fuzed projectiles.

Interestingly, these fuzes were called variable time, or VT, fuzes in order to mask their true nature, this nomenclature has carried over to this day. Guided missiles have a higher incidence of direct hits than that of unguided projectiles, but the original principles still hold true.

Proximity fuzes accomplish their purpose through "influence sensing," with no contact between the warhead and the target. These fuzes are actuated by some characteristic feature of the target rather than by physical contact with it. Initiation can be caused by a reflected radio or electro-optical signal, an induced magnetic field, a pressure measurement, or an acoustical impulse.

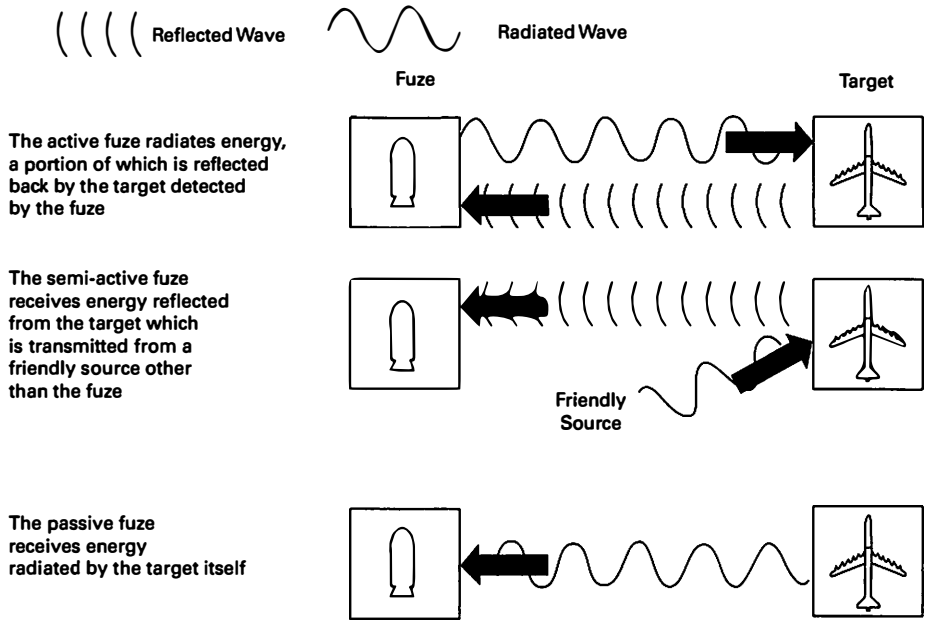


FIGURE 16-5 *Fuze classification by mode of operation*

A proximity fuze is classified by its mode of operation, of which there are three: active, semiactive, and passive. These three modes are illustrated in figure 16-5.

Electromagnetic Proximity Fuzing

Conceivably, all portions of the electromagnetic spectrum could be used for target detection. Practically, however, considerations of propagation, attenuation, and other parameters affected by radiation determine the applicability. The portions of the spectrum having greatest utility are radio, radar (microwaves), and infrared.

An electromagnetic fuze, operating particularly in the radio and radar regions, may be constructed to operate much like a miniature radar set. It must transmit, receive, and identify electromagnetic pulses. The correct received signal initiates the detonator. The basic active electromagnetic proximity TDD has the following components:

1. A transceiver that is capable of delivering the required power for transmission and sensitive enough to sense the weak signal return
2. Amplifying circuitry to magnify the return signal, so that it will activate the firing circuit and initiate the detonator. The receiver and amplifier circuits are designed to select the correct signal
3. A power supply to generate and provide electrical power for the fuze

Some weapons used against surface targets, such as bomblets delivered to the target area in canisters called cluster bomb units (CBU) and fuel-air-explosive (FAE) weapons, employ proximity fuzes to deploy and disperse the payload at a predetermined height.

Antipersonnel weapons having unitary warheads are more effective when detonated above the target area rather than on contact. Proximity fuzes for these applications may function as radio altimeters or as slant-range-sensing devices that measure range to the surface at the projected point of weapon impact.

One means of signal selection makes use of the radar principle, in which the elapsed time between a transmitted and received pulse is a function of the range between the target and the weapon. A range-gate circuit set for a given distance will pass the signal to initiate the warhead when the elapsed time reduces to a predetermined value. The maximum range gate, which renders the fuze insensitive to effects that are farther away in range, is called the range cutoff of the TDD sensor; see figure 16-6.

For example, with respect to backscatter from precipitation, the magnitude of fuze response is proportional to the volume of rainfall interrogated within an "integration period." When this volume is restricted by the range cutoff, fuze response to rain return is reduced by an amount proportional to the reduction in volume of rain interrogated.

This mechanism is especially important in the engagement of low-altitude targets, where the sea surface is a highly efficient reflector of radar waves.

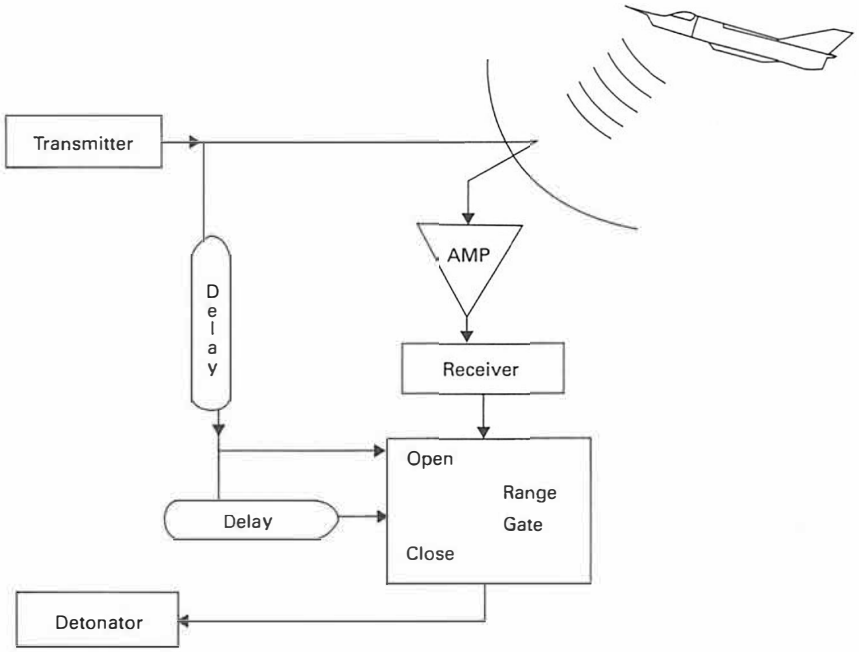


FIGURE 16-6 *Electromagnetic fuze circuitry*

Another means of signal selection uses the Doppler principle, in which the frequency of the received signal varies as a function of the relative velocity between the weapon and the target. This permits classification of the targets according to their radial velocities, which is useful in the selection of a primary target within a group of signals from a variety of sources.

The Doppler frequency can also be used to determine when to detonate the warhead. If the encounter is head-on, the Doppler would be a relatively high upshift, in comparison to predetermined levels set within the fuze circuitry, and detonation of the warhead would need to be immediate to ensure hitting the target and not going behind it.

If the encounter is from the rear, as in a chasing situation, the Doppler shift would be relatively low, and therefore a delay in detonation might be desired to ensure a hit. The point of detonation could occur when the Doppler shifts from an up to a down Doppler, which is at the point of closest approach.

In surface-to-air and air-to-air weapon applications, the main function of the proximity fuze is to compensate for terminal errors in weapon trajectory by detonating the warhead at a point calculated to inflict a maximum level of damage on the target. Because of kinematic relationships among the weapon, the target, and the warhead kill mechanism following detonation, the preferred point of detonation is not, in general, the point of closest approach of the weapon to the target.

Instead, the weapon engineer strives for a fuze design that will adaptively initiate the warhead within the so-called "lethal burst interval" for each trajectory that the weapon is likely to experience in operational usage. The lethal burst interval is defined as "that interval along the trajectory in which the warhead can be denoted so as to hit the target in the vulnerable area."

That ideal state is, of course, never realized in practice, here idealism must be tempered by practicality. As a compromise, the engineer derives a design that is practical and affordable, and that will maximize the number of lethal detonations in a representative sample of all postulated encounters.

The actual point of detonation realized in a given encounter is the result of fuze design parameters, weapon vector velocity, location of the encounter in the weapon performance envelope, burst control logic, and a wide

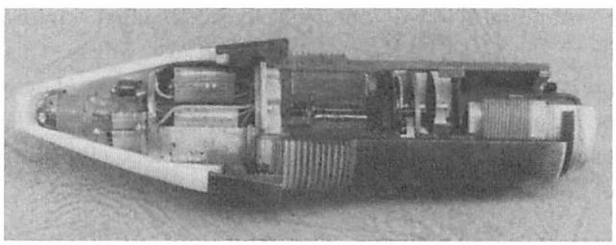


FIGURE 16-7 *An RF proximity fuze for a gun projectile*

variety of other contributing factors such as active or passive countermeasures and environmental effects.

To understand this, it will help to realize that tactical missile warheads are not so powerful that they will devastate the target if detonated just anywhere in its vicinity. Also, warheads cannot be made larger and heavier without seriously compromising missile speed, altitude, range, and maneuverability. Therefore, the fragments ejected by the warhead must be delivered onto those areas of the target that are most vulnerable to damage: the cockpit, the engine compressor stages, and the flight control systems. This requires sophisticated fuzing, with precise burst control.

Furthermore, to maximize its lethal range, the warhead is designed so that fragments are confined to a narrow zone in the missile pitch plane. Thus, a lethal density of fragments is maintained to ranges much greater than if the fragments were dispersed isotropically.

The proximity fuze is called upon to solve this convergence problem of converging the narrow fragment pattern and the limited vulnerable regions of the target. Figure 16-8 illustrates how this is accomplished. The figure represents a missile approaching the target from the rear and passing either above or below the target—not a direct hit. Target vulnerable areas or regions that must be hit to ensure a kill are also depicted.

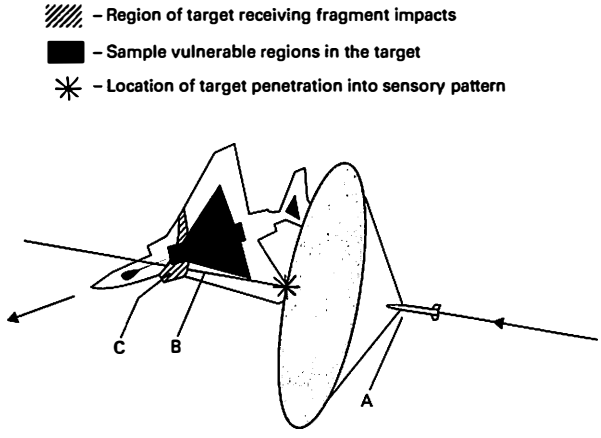


FIGURE 16-8 *Representative terminal engagement situation*

At point *A*, the proximity fuze detects the first extremity of the target that penetrates its sensory pattern. This initiates a series of events that will result in warhead detonation at point *B*. The time delay that permits the missile to traverse from point *A* to point *B* is varied as a function of closing velocity in accordance with the equation

$$T = \frac{M}{V_c} - N \qquad (16\text{-}1)$$

where:

V_c = closing velocity (m/s)
T = time (seconds)
M and N = either constants or variables, depending on the complexity of the fuze, with their units chosen accordingly

When the warhead is detonated at point *B*, the vector sum of missile velocity, warhead expansion velocity, and target velocity will cause the warhead fragments to impact the target in region *C*. This illustrates a successful warhead detonation, as the fragments impact some of the vulnerable regions of the target.

It should be evident immediately that the terminal encounter situation involves the interplay of many parameters that depend upon target characteristics, missile characteristics, and positional relationships at the time of missile launch. The degree of variability and uncertainty in these parameters is increased by the maneuverability and countermeasure capability of the target.

Because most missiles are capable of an all-aspect attack, and because the target's velocity can vary over a wide range of values, the closing velocity used in the time-delay equation may range from 180 to 1,800 meters per second, or more. The spectrum of targets intended for the missile will encompass those that may be small, medium, or large in physical size but whose velocity capabilities overlap to such an extent that a time delay selected for a given V_c might be successful for one target but a failure for another.

Furthermore, it is possible for very long targets in high-crossing-angle encounters to present only a square meter or less of vulnerable area to the warhead. Thus, gross target dimensions are not very useful as measures of where the warhead should best be detonated.

Because they are sensor devices by nature, fuzes are subject to a wide spectrum of disorienting influences in the real-world tactical environment. One of the foremost prerequisites of good fuze design, therefore, is to devise a sensor system that discriminates bona fide target return from all distracting influences, whether they take the form of electronic or optical countermeasures, intense electromagnetic radiation levels characteristic of fleet environments, chaff, precipitation, out-of-range targets, or the radar clutter that occurs in missile flight at low altitude.

Electro-Optical Applications

Electro-optical devices such as lasers provide another means of detecting the target. An active optical target detection device normally consists of laser-emitting

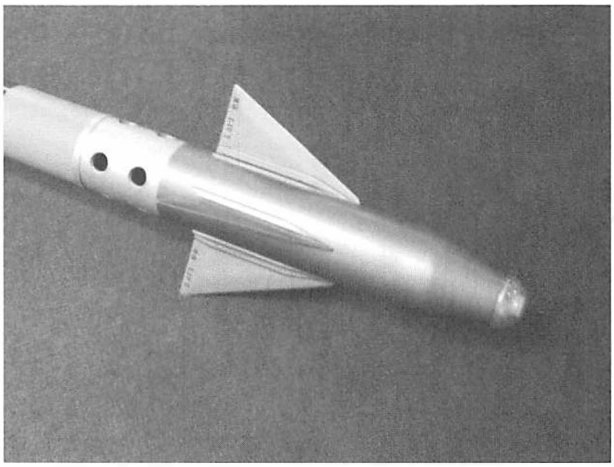


FIGURE 16-9 *Sidewinder missile with laser TDD*

diodes and light-sensor diodes arranged in such a manner that they create a spoke-like pattern in front of and around the missile. If the sensor diodes receive a return of the laser and a predetermined threshold is met, the control system will recognize that the missile is next to the target and trigger the warhead. Figure 16-9 shows the laser-sensor pair immediately behind the canard control fins.[i]

Magnetostatic Fuze Applications

Magnetic sensors such as magnetic anomaly detection (MAD) measure changes in the earth's magnetic field or the presence of a source of magnetic flux. In the case of fuze systems, a magnetic sensor is designed to recognize an anomaly and ultimately cause the fuze to function.

Such a target detection device is designed to close a firing circuit upon a disturbance of the earth's magnetic field by one of the magnetic components of a ship. The target does not necessarily have to be moving, although a type of fuze may be employed that requires a certain rate of target movement. Underwater mines that employ this principle of fuzing are commonly used. They are known as dip-needle, influence mines.

Another type of magnetic TDD employed with mines is an inductor mechanism, frequently called a search coil. This device employs the principle that a magnetic field induces current flow in a conductor as the field changes relative to the conductor.

The small voltage induced in the coil is amplified and then caused to energize the firing circuit, which in turn initiates the detonator. The extreme simplicity of this device makes it highly reliable and difficult to counter.

Magnetostatic fuzing is also used for subsurface targets. The magnetic field disturbance fuze for subsurface targets is also actuated by a change in the surrounding magnetic field. Any change in the magnitude of the magnetic field activates the fuze.

A magnetic-type fuze provides the possibility of damaging a target without a direct hit. This is important, as the damage potential is greater when the explosion of the warhead takes place several feet below the hull, rather than at the side or near the surface of the water, as discussed in chapter 18.

A very advanced method of fuze initiation caused by a ship's magnetic field is employed in a torpedo electromagnetic detecting fuze. Such a system operates on what can be called the generator principle. Essentially, an electric generator consists of a coil of wire rotated in a magnetic field to produce a voltage. Similarly, a small voltage is developed across a coil of wire (the search coil) when it comes in contact with a moving or changing magnetic field.

However, a complex problem occurs, due to the fact that as the torpedo moves through the water, it creates its own change in the field gradient, which might initiate the fuze at an earlier time than intended. This has led to the development of the gradiometer, a device attached to the torpedo that has two search coils approximately 1 foot apart that are connected in opposing series. As the now magnetically balanced torpedo moves in the earth's magnetic field, equal and opposite voltages are induced in the coils, and no net voltage results.

In the vicinity of a steel vessel, the situation is different. One of the two coils is slightly closer to the vessel than the other, and a slightly different voltage will therefore be induced in it. This difference is small, but when properly amplified, it causes the detonator to explode the warhead.

Acoustic Fuze Applications

Acoustic disturbances, such as propeller and machinery noises or hull vibrations, invariably accompany the passage of a ship through the water. The intensity or strength of the sound wave generated depends upon several factors such as ship size, shape, number of propellers, and type of machinery.

Therefore, a ship's acoustic signal is variable, and acoustic fuzes must be designed to prevent an intense signal from actuating the fuze at distances well beyond the effective explosive radius of the payload. Figure 16-10 illustrates a basic acoustic mine mechanism.

An acoustic fuze on a mine-type system employs a hydrophone as a detector to sense the presence of a target. A typical hydrophone functions much the same as the human ear. A diaphragm, corresponding to an eardrum, vibrates from the impact of underwater sound

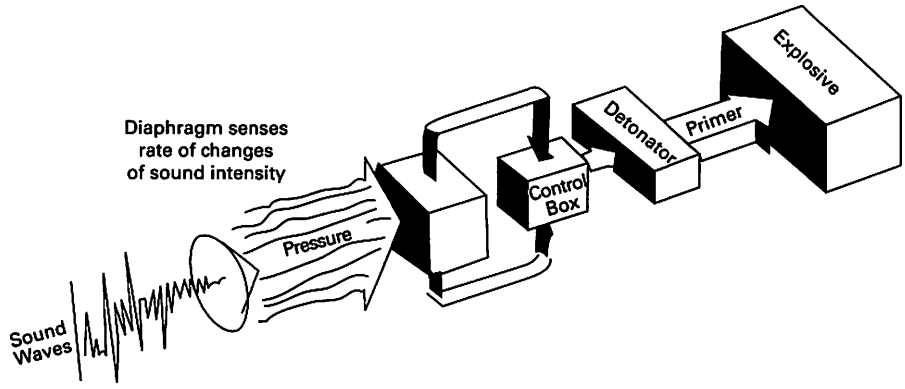


FIGURE 16-10 *Acoustic mine mechanism*

waves. These vibrations are transmitted through an oil medium to a crystal, which converts the mechanical energy to the electrical energy required to initiate the firing mechanism.

The selectivity of the firing mechanism is so critical that only the pulses of the required characteristics are sent to the detonator. Selectivity is necessary because of the varied sounds that are received by the weapon. Acoustic devices are designed to be unresponsive to sounds representing nontarget objects such as a passing school of fish.

Another design requirement is preventing these fuzes from being easily actuated by countermining noises such as nearby explosions. Acoustic proximity fuzes are usually given a predetermined sound intensity. If the received sound builds up too quickly, as illustrated in figure 16-11, or too slowly, the discriminating features incorporated in the fuze circuitry will not allow the fuze to operate.

Therefore, acoustic firing mechanisms possess a very selective type of hearing and processing to determine valid targets from other noise sources such as countermeasures. For example, a 4,000-ton tramp steamer may have one large propeller turning over rather slowly, and a 4,000-ton destroyer may have two propellers turning over much faster. An acoustic firing mechanism must distinguish between the steamer and the destroyer, and fire on the selected target. When the firing mechanism detects a sound that has the required characteristics, including intensity and rate of change of intensity, the mechanism initiates the firing circuitry and fires the detonator.

Acoustic fuze mechanisms are used in torpedoes as well as mines. There are two operating modes of acoustic torpedoes, the active and the passive types. The passive type has a homing device that guides the torpedo in the direction of the strongest target noise.

The active type employs a sonar set that emits a series of sonic pings that are reflected back to a receiver in the torpedo. The principle is similar to that of a sonar set. As the torpedo approaches the target, less time is required for a

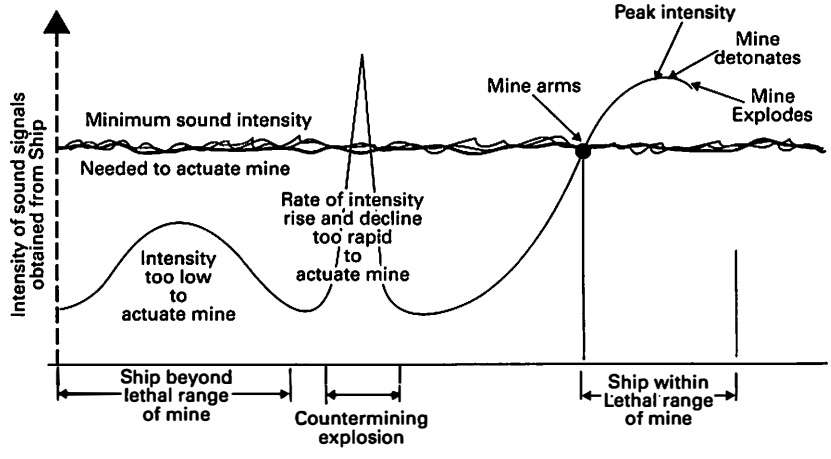


FIGURE 16-11 *Sound variation due to a ship's travel*

signal to travel to the target and return. At a predetermined critical distance, initiation of the firing circuit begins.

Seismic Fuzing

A similar type of acoustic influence sensor used in some types of mines is the "seismic" firing mechanism. This sensor is essentially an acoustic fuze, but it receives its threshold signal in a lower bandwidth through weapon case vibration. These sensors can be made extremely sensitive and may provide for both land-based and in-water application.

They offer advantages over the pure acoustic mine fuze in that they may be set over a wider range of selectivity and may incorporate special countermeasure features. Most new naval mines will use a seismic fuze in combination with other influence fuzes as a means of ensuring weapon activation on a valid target and of reducing the effectiveness of minesweeping efforts.

Hydrostatic (Pressure) Fuzing

Ocean swells and surface waves produce pressure variations of considerable magnitude. Moving ships displace water at a finite rate; see figure 16-12. This continuous water flow is measurable at considerable distances from the ship as pressure variations that normally exist in the water. Various pressure-measuring mechanisms can be used in fuzes to detect such variations.

The pressure differential becomes more pronounced when the ship is moving through confined waters, but is still appreciable in the open sea, even at a considerable depth. This pressure variation, called the pressure signature of a ship, is a function of ship speed and displacement and water depth.

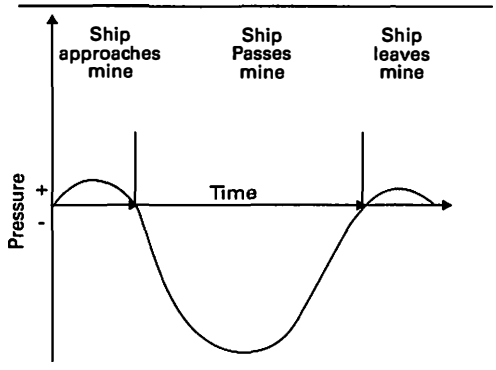


FIGURE 16-12 *Pressure variation due to a ship's travel*

To avoid premature firing because of wave action, pressure-firing mechanisms are designed so as to be unaffected by rapid pressure fluctuations. Pressure sensors are commonly associated with bottom mines and are extremely difficult to counter through normal influence-mine countermeasure techniques. Pressure-firing mechanisms are seldom used alone, but are generally combined with other influence firing devices.

Underwater Electric Potential (UEP)

Another type of proximity fuze used in naval mines senses the influence of a ship's underwater electric potential field. A variety of shipboard sources cause electric currents in water that are detectable. These sources include galvanic potential of metals in sea water, the active and passive corrosion protection systems employed by ships, ships' power systems, and stray electromagnetic fields produced by electrical equipment. Though each effect is small, each is detectable and can be distinguished from the environment.

Combination Fuzing

Systems involving a combination of influences are available in most mine firing devices. The combinations of magnetic, pressure, and acoustic/seismic systems are used to compensate for the disadvantages of one system with the advantages of another. Mine countermeasure effectiveness is greatly reduced through use of combination fuzing.

Mine Performance Data

Minefield theory involves detailed and complex analysis, which is beyond the scope of this text. The entire minefield must be considered as a weapon in a mining scenario, and multiple targets are programmed against it to determine the overall effectiveness. For the sake of simplicity, our discussion will be restricted to a single mine against a single target. The influence sensors described earlier may be considered individually or as a combination.

Regardless of the particular influence, there is a threshold level that will activate the mine based on the sensitivity setting for the weapon. This will obviously occur at some distance from the target—the greater the range, the more sensitive the mine setting—and is called the actuation distance, as depicted in figure 16-13.

It should be obvious that a mine could be set so sensitively that it could be made to detonate on a target influence such that the explosion would be outside of the range at which the mine could do damage to the target.

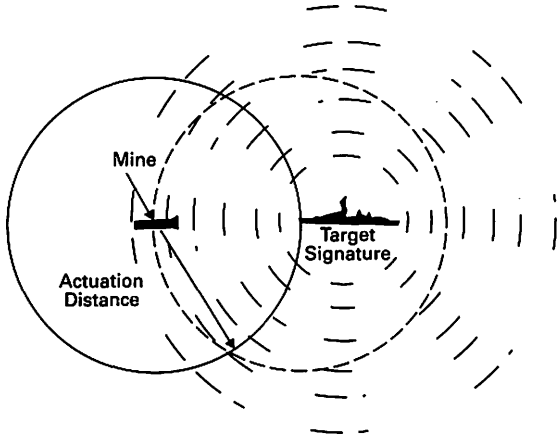


FIGURE 16-13 *Mine actuation distance*

The range at which a mine will damage a target is referred to as the damage distance. This would also be described as similar to the circular pattern in figure 16-13.

From a minefield planning standpoint, it is most important to optimize the mine fuze setting such that the weapon does not actuate until the target is within damage range. Obviously, the actuation distance should be less than or equal to the mine damage distance. The problem becomes much more complex in actual planning. Each target class has its own associated signature as well as its own relative toughness against damage, and the ship-mine encounter is greatly influenced by such factors as ship heading and speed, environmental state, and orientation of mine and ship.

It is impossible to determine all of these factors precisely; therefore, approximations are made at varying levels of accuracy. Mine actuation probabilities and damage probabilities are computed based on various target and mine types, water depths, ship speeds, mine sensitivities, and so on, and are published as classified information for use in minefield planning methodology.

Nonproximity-Type Fuzes

Time Fuzes

A time fuze uses a timing device that causes warhead detonation at some predetermined time. This device, or "clock" mechanism, is actuated by a signal from a fire control system that provides the fuze with the time of flight necessary for the weapon to reach the predicted future position of the target or vice versa. The clock mechanism can thus detonate the payload at the exact moment necessary to cause target damage.

Point Detonating Fuzes

Point detonating, impact, and contact fuze are common names for a fuze that must physically touch the target to cause the warhead to detonate. Detonation can be made to occur just outside or just inside the target, depending upon the design of the fuze. The means of actuation may be through the acceleration of a mass, the joining of two points, or the removal or rupture of a barrier.

When properly designed, an impact fuze can be used to produce the detonation of a bursting charge at any point from the surface of the target to a measured distance within the target. The electrical or mechanical action of impact fuzes is usually actuated by moving a firing pin, completing an electrical circuit, or having a piezoelectric transducer initiate an electronic detonator. Figure 16-14 shows a typical projectile impact fuzing system.

Fuzes initiated by impact with the target generally afford the simplest solutions for fuzing; however, the design of such fuzes is by no means simple. Difficulty arises in the design of impact-type fuzes for targets that are not simple planer surfaces perpendicular to the line of flight and are not extremely solid, such as an airplane. The "skin" of an airplane generally is so thin that it does not decelerate the projectile enough as it penetrates to activate the impact element.

This is particularly true in the case of high-speed missiles for which the impact element must be designed to withstand very high speeds. Additionally, because of irregular target shapes, warheads may strike at an acute angle, rather than perpendicular to the target, causing the warhead to glance off without sufficient activating force.

If the target is a solid surface such as a ship or submarine, the above problems are of no consequence; however, because of the large impact forces, functional reliability of the fuze itself does become an issue.

Naval mines provide a glimpse of the different applications of the contact fuze concept. Naval mine contact fuze types include the electrochemical horn, the switch horn, and the galvanic fuze. Figure 16-15 is an example of an electrochemical horn. Here, when contact is made of sufficient force, a glass vial of acid is broken that then comes into contact with zinc and carbon elements, creating an electric current that is used to initiate detonation.

The switch horn, figure 16-16, is a simple lever that closes a detonation circuit.

A galvanic fuze acts like a fuze and sensing element combined, because it is displaced from the warhead. This fuze uses the sea water as an electrolyte, with a small copper float attached to a long wire, as shown in figure 16-17.

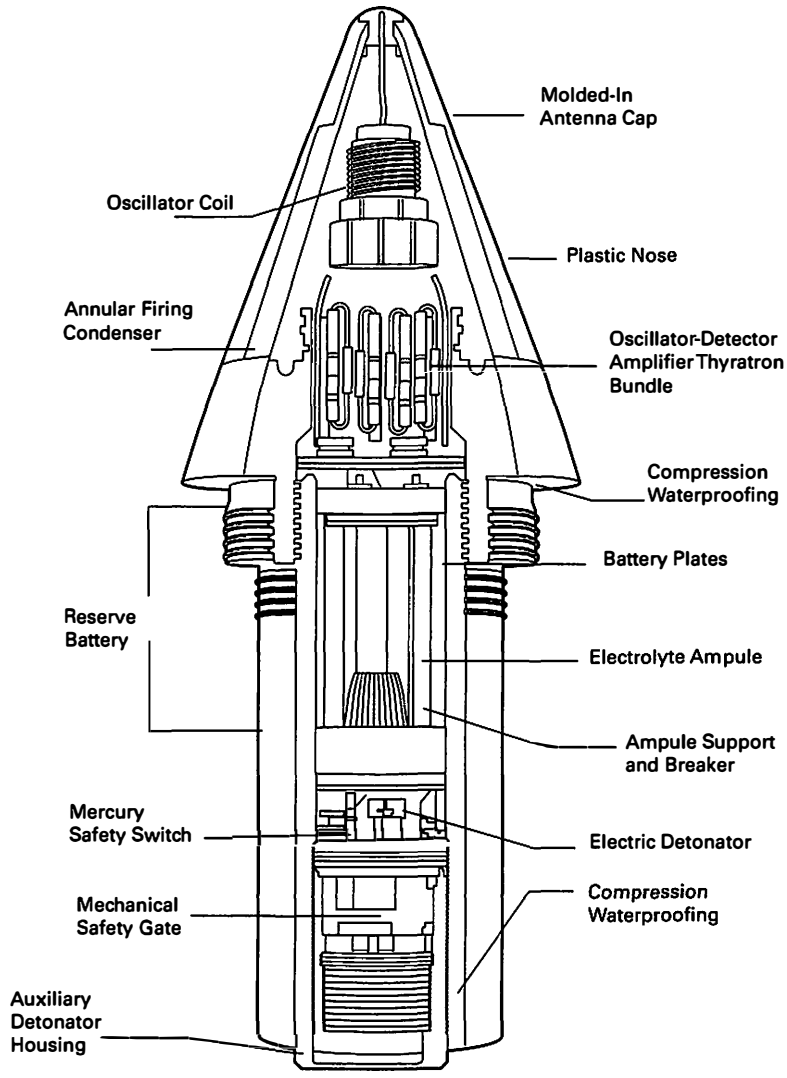


FIGURE 16-14 *Projectile impact fuze*

When the steel hull of a ship touches the float or the wire, a current is produced that then activates the mine.

Delay Fuzes

Delay-action-type fuzes are primarily used for armor piercing or ground penetrating. These fuzes initiate detonation at some measurable time after impact. One method of delaying the initiating action of the fuze is to use delay pellets, which burn for some predetermined time. After this slow burn time, the detonator explodes, starting the explosive train.

Another method is to allow the fuze to sense the deceleration of the warhead and start the detonation sometime after maximum deceleration. Still another method uses air pressure to sense voids in structures as the warhead penetrates various layers of the target.

Regardless of how the fuze achieves its delay, these fuzes must be made robust enough to function reliably after impact. The location of the fuze unit is selected to protect the fuze during penetration of the target and to provide a higher probability that the fuze will function as intended.

Command Fuze

A command fuze is a device that receives the detonation command from a source not located on the weapon itself. This type of fuze waits until a signal from the weapon control station is detected before it starts the detonation sequence.

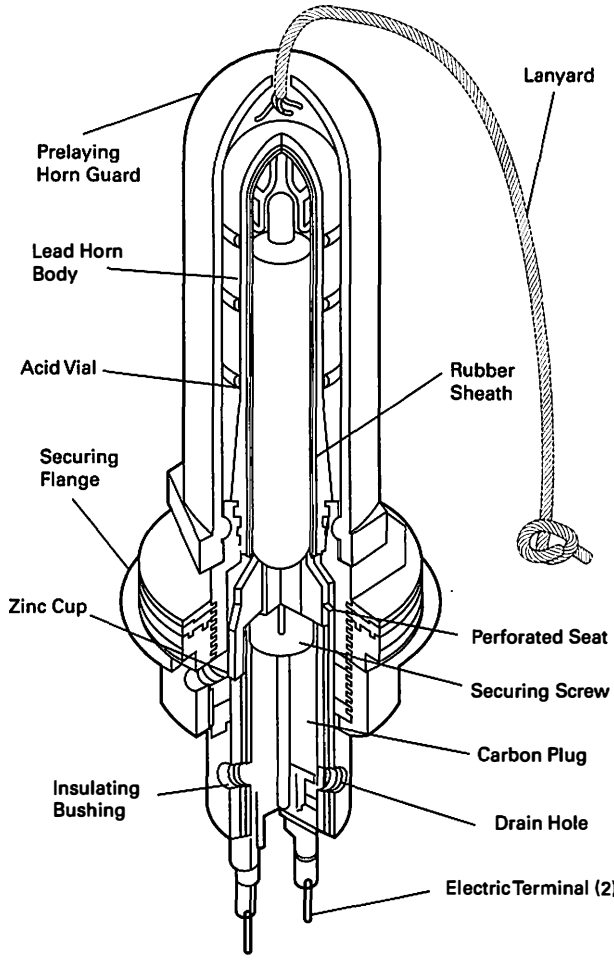


FIGURE 16-15 *Impact mine fuze*

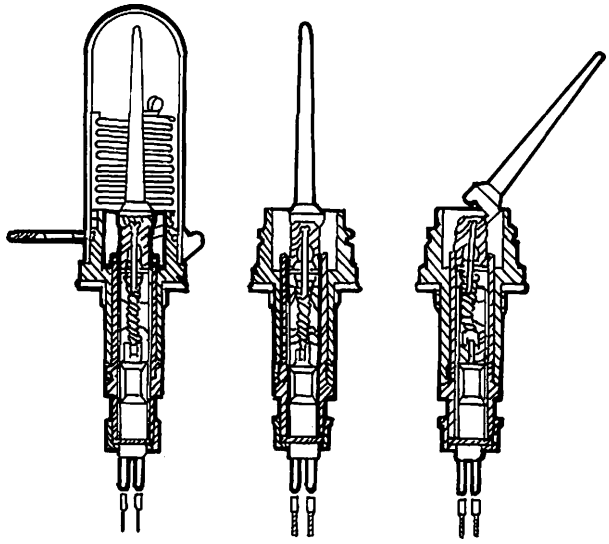


FIGURE 16-16 *Simple switch-type fuze*

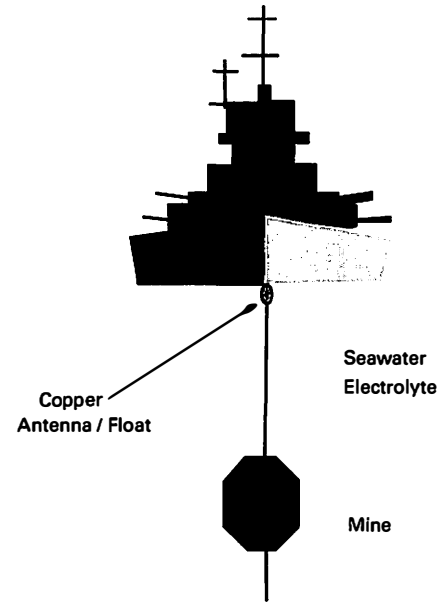


FIGURE 16-17 *Galvanic-type mine fuze system*

Reliability

The reliability functions—target sensing, arming, and payload initiation—are obviously contrary to the weapon safety function of a fuze. A fuze detonator that will reliably detonate the payload at target impact presents a safety reliability problem if it also detonates when dropped during handling.

To design a fuze to be reliably safe, a designer may compromise arming and firing reliability. A fuze detonator that is designed not to detonate from rough handling may also be so insensitive that the contact fuze would dud at target impact. An S&A device that is designed to sense too high an airspeed may not arm in time when deployed against short-range targets.

There are several methods available to the engineer to increase arming and initiation reliability and also safety reliability without compromising the reliability of each of these paradoxical functions.

Functional Reliability

Although fuze design must emphasize safety, the primary objective of the warhead is to detonate at the proper instant. Therefore, the sequence of safing and arming and target-sensing devices must have a high probability of functioning correctly when they receive their unique inputs. Functional failure of any one of the devices shown in figure 16-18 would result in a dud warhead.

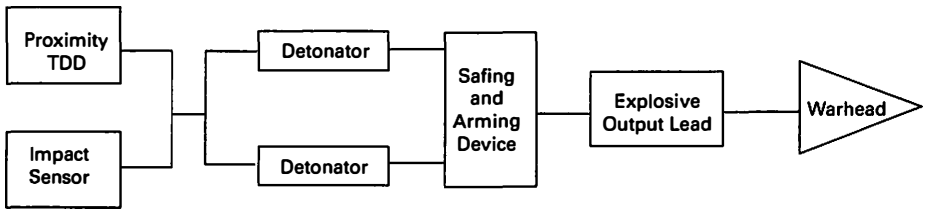


FIGURE 16-18 *Fuze system redundancy*

As with any man-made device, the fuze components must be based on a sound design while avoiding overly complex functional requirements. Before the reliability criteria can be ensured, every new S&A device must be extensively tested.

Sophisticated weapons have devices that sense and arm with such a high reliability that fewer than 1 in 1,000 is expected to fail functionally. Note that a functional failure is not the same as a safety failure, but occurs when the component fails to operate as designed when proper environmental conditions exist.

Functional reliability of the entire fuze system can be increased by having more than one type of target sensor placed in parallel and more than one detonator in the S&A device, as shown in figure 16-18. Most missile fuze systems have one or more backup contact sensors in addition to their proximity sensor. Thus, the probability of warhead functioning is increased.

This leads to a basic principle in fuze design, which can be stated as:

> A redundant increase of the number of similar components placed in parallel in the firing path will increase the arming and firing reliability.

Safety Reliability

As pointed out earlier in the safing and arming discussion, safety reliability depends on the successful sequential operation of redundant components. A safety failure occurs when the components, for some reason, fail to keep the fuze in a safe condition when so desired and premature arming occurs. Thus, a safety failure is said to have occurred if the weapon becomes armed other than through its normal firing sequence.

From this definition, it can be seen that for a safety failure to occur, the series components of a safing and arming device must simultaneously malfunction in such a manner that the weapon is armed. Since the probability that a component of an S&A device will have a safety failure is normally very small, the probability of three or four components having a safety failure simultaneously is minute.

Failure of any component in the series results in overall failure to reach an "armed" status. Thus, a similar fundamental principle in fuze design can be stated as:

> A redundant increase in the number and types of devices placed in series in the firing path increases the safety reliability of a fuze.

Probability Theory

The performance of complex systems such as weapon fuzes can be determined through the application of probability theory. The principles of parallel and series operations of components as stated above lead to a method of determination of total system reliability regardless of the number or arrangement of components. Two fundamental principles of probability theory are used in determining fuze reliability: statistically independent events and mutually exclusive events.

Statistically Independent Events

Two events, A and B, are independent if the probability that A occurs is unaffected by the occurrence or nonoccurrence of B, and vice versa. The probability of several statistically independent events occurring simultaneously is the product of the probabilities that the individual events occur. Thus, the probability that events A and B both occur, given that their probabilities of occurrence are $P(A)$ and $P(B)$ is:

$$P(A \text{ and } B) = P(A) \times P(B) \qquad (16\text{-}2)$$

Mutually Exclusive Events

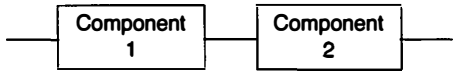
The probability of several mutually exclusive events occurring is the sum of the probabilities of each event. Two events are said to be mutually exclusive if the occurrence of one event prevents the occurrence of the other. If events A and B are mutually exclusive and if their probabilities of occurrence are $P(A)$ and $P(B)$, then the probability that either A or B will occur is:

$$P(A \text{ or } B) = P(A) + P(B) \qquad (16\text{-}3)$$

Mutually exclusive events are not independent since, by definition, if one occurs, the other cannot, or $P(A \text{ and } B) = 0$.

Series Systems

Suppose that a system consists of two independent components arranged in series. That is, for the system to operate properly or be successful, each of these independent components must operate properly and in sequence. Schematically this system can be depicted as follows:



The effect of each component's success, *S*, or failure, *F*, on total system performance can be tabulated as follows:

Component 1	*S*	*S*	*F*	*F*
Component 2	*S*	*F*	*S*	*F*
System	*S*	*F*	*F*	*F*

Notice that the system is successful only if both of the components are successful. All other possible situations lead to system failure. In order to calculate the probability of this system's success, it is necessary only to calculate the probability that both components are successful.

Letting S_1 and S_2 stand for the probability that component 1 and component 2 are successful, respectively, and noting that these components are independent, a simple expression for the probability of system success, $P(S)$, results:

$$P(S) = S_1 \times S_2 \tag{16-4}$$

Also, since the qualities of success and failure are mutually exclusive and cover all possible system outcomes:

$$P(S) + P(F) = 1 \tag{16-5}$$

or

$$P(F) = 1 - P(S) = 1 - (S_1 \times S_2)$$

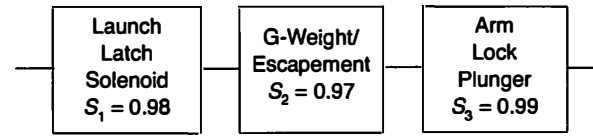
If a system has more than two components in series, its probability of success is again simply the product of the individual component success probabilities:

$$P(S) = S_1 \times S_2 \times S_3 \times \cdots \tag{16-6}$$

and

$$P(F) = 1 - P(S) = 1 - (S_1 \times S_2 \times S_3 \times \cdots)$$

EXAMPLE 16-1: The S&A device schematically shown below has three independent components that must operate in series for arming to occur. The system can be modeled as follows:



SOLUTION: The probability of successful arming for this device can be calculated as

$$P(S) = S_1 \times S_2 \times S_3 = 0.98 \times 0.97 \times 0.99 = 0.9411$$

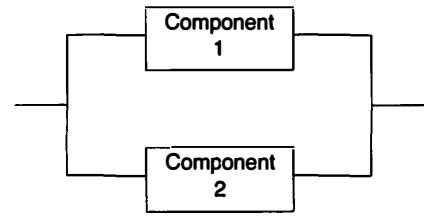
Also, the probability of failure is

$$P(F) = 1 - P(S) = 1 - 0.9411 = 0.0589$$

Note that calculations are usually carried out to four significant places and that the resulting probability (in series systems) is less than any individual original probability.

Parallel Systems

Suppose that a system consists of two independent components arranged in parallel. This means that for the system to be successful, only one of the components must be successful. Schematically, this system can be depicted as follows:



The effect of each component's success, *S*, or failure, *F*, is illustrated in the following table:

Component 1	*S*	*S*	*F*	*F*
Component 2	*S*	*F*	*S*	*F*
System	*S*	*S*	*S*	*F*

Notice that for a parallel system, the only way that the system can fail is if all the individual components fail simultaneously. Calculating the probability of system success is rather complicated since it is the probability that component 1 and component 2 are successful or that component 1 is successful and component 2 fails or that component 1 fails and component 2 succeeds. It is much easier to calculate the probability that the system fails, which is:

$$P(F) = F_1 \times F_2 \tag{16-7}$$

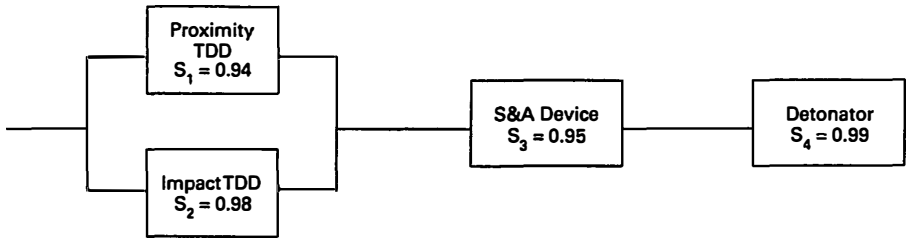
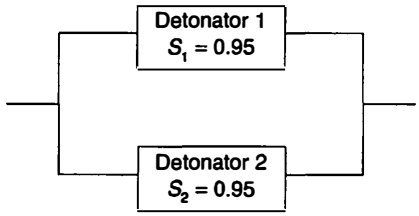


FIGURE 16-19 *Fuze system reliability example*

Then, since system success and failure are mutually exclusive,

$$P(S) = 1 - P(F) = 1 - (F_1 \times F_2) \qquad (16\text{-}8)$$

EXAMPLE 16-2: Consider a system consisting of two detonators placed in parallel, each with a success probability of 0.95.



SOLUTION: The probability of success for this system is given by:

$$P(S) = 1 - (F_1 \times F_2)$$

Now, since the qualities of success or failure for each individual component are mutually exclusive:

$$S + F = 1$$

or

$$F = 1 - S$$

Thus,

$$\begin{aligned} P(S) &= 1 - (F_1 \times F_2) = 1 - [(1 - S_1) \times (1 - S_2)] \\ &= 1 - [(1 - 0.95) \times (1 - 0.95)] \\ &= 1 - (0.05 \times 0.05) \\ &= 1 - 0.0025 \\ &= 0.9975 \end{aligned}$$

EXAMPLE 16-3: Fuze system reliability. A particular type of ammunition used for shore bombardment against troops in the open is designated as VT-NSD (variable time–non-self-destruct). The fuze system for this type of projectile has a proximity sensor designed to detonate 100 feet from the ground and an impact sensor as a back-up, should the proximity sensor fail. The diagram in figure 16-19 is a schematic representation of the entire fuze system, and each component's probability of operating successfully is given.

SOLUTION: To calculate the fuze system reliability (probability of success), it should first be noted that this is a combination of parallel and series elements. If the probability of the TDD function success is first calculated, then this system is reduced to three functional elements in series. Calculating the probability of the TDD function success,

$$\begin{aligned} P_{TDD}(S) &= 1 - P_{TDD}(F) = 1 - (F_1 \times F_2) \\ &= 1 - [(1 - S_1) \times (1 - S_2)] \\ &= 1 - [(1 - 0.94) \times (1 - 0.98)] = 0.9988 \end{aligned}$$

Now, calculating the overall success probability of three functional elements in series:

$$\begin{aligned} P(S) &= P_{TDD}(S) \times S_3 \times S_4 \\ &= 0.9988 \times 0.95 \times 0.99 \\ &= 0.9394 \end{aligned}$$

Thus, this notional fuze system will successfully detonate the warhead approximately 94% of the time.

[i] Harris

17 Chemical Explosives

Introduction

An explosion results in the rapid and violent release of energy and may be divided into three general types: mechanical, chemical, and nuclear. A mechanical explosion, such as the buildup of pressure associated with a steam line rupture or a boiler explosion, is of little military application and will not be discussed here. This chapter will concentrate on chemical-type explosions. Nuclear explosions will be discussed in chapter 19.

An explosive may be defined as a material (chemical or nuclear) that can be initiated—meaning it is not spontaneous but controlled—and undergoes very rapid, self-propagating decomposition, resulting in:

1. Formation of more stable materials and gases
2. Liberation of heat
3. Development of a sudden pressure effect through the action of heat on the gases produced or in situ gases

A military warhead is designed to inflict damage on a target through the rapid transfer of energy from the warhead to the target. This energy is typically mechanical in nature, manifesting itself in the form of a shock wave, fragments from the warhead casing, or both. In order to be effective, a large amount of energy must be released in an extremely short amount of time, and for most warheads that energy is stored within the chemical explosive.

Categories of Explosives

Not only must explosive materials be highly energetic, but they must also react violently. The speed of the reaction is vital to the buildup of a large amount of energy within what is initially a small volume. Reactions may proceed so slowly that the released energy dissipates, a consideration to keep in mind if one does not want the generated shock wave to interact with the target.

An explosion, however, creates a shock wave capable of shattering metal casings and ejecting fragments outward from the point of the explosion, or of generating a pressure wave of such magnitude that damage effects are felt as the blast wave passes over or through the target near the detonation point. If the energy release is slow, the shock wave will be gradual and extended, and the initial fragment velocities will be low. Violent, explosive reactions, however, produce very sharp, short-duration, high-pressure shock waves and high-fragment velocities.

The property of the shock wave is related not only to the explosive material, but also to the initial degree of confinement. If an explosion is initially restrained such as in a grenade or bomb casing, large pressures may build and yield the same effects produced by a significantly larger portion of explosive material in which the reaction is not initially confined. This rapidity of reaction, or rate of decomposition, is of such importance that it is used to classify explosive materials.

High Explosives

Explosive materials that react very violently, producing almost instantaneous decomposition, are known as high explosives. This extremely rapid chemical transformation is referred to as detonation, and as one might expect, high explosives are used solely for their destructive power as the main explosive charges in warheads. High explosives may be pure compounds or mixtures of several compounds with additives—for example, powdered metal such as aluminum, plasticizing oils, or wax—that impart desired stability and performance characteristics.

The detonating high explosive produces a shock wave that travels through the explosive material at supersonic speeds, ranging from 1,000 meters per second to 8,500 meters per second. High explosives are further subdivided

by the nature of their sensitivity to shock, friction, heat, and electric charge.

Primary high explosives This type of explosive is extremely sensitive to shock, friction, heat, or spark and will burn rapidly or detonate if ignited. Primary high explosives are frequently used as the first element—a primer or detonator—in an explosive series to provide the motive force to set off the booster and main (bursting) charge.

Secondary high explosives These explosives are relatively insensitive to shock, friction, and heat, requiring a significant force to start the reaction. Secondary explosives are frequently used as the booster charge and the main (bursting) charge of a warhead.

Low Explosives

In contrast to primary high explosives, some chemical materials react more slowly in their decomposition, burning in a more controlled fashion, and are known as low explosives. The relatively slower decomposition of the explosive material is known as deflagration, and produces subsonic shock waves with velocities reaching 400 meters per second through the explosive material.

This process still releases a large amount of energy, but due to its relatively slow rate of reaction, the energy is more usefully channeled as a propellant charge where the expanding gases are used to move projectiles. Examples include black powder and smokeless powder.

As stated earlier in this chapter, the nature of the shock wave associated with decomposition of the explosive substance is partially determined by the nature of the confinement of the explosive material. Therefore, it is entirely possible, given the right circumstances, to change a low explosive from a deflagrating substance to one that detonates—for example, gasoline—and to change a high explosive from detonating to one that deflagrates.

The Detonation Sequence and the Explosive Train

As previously stated, primary and secondary high explosives are used in the detonation sequence for a typical warhead. This arrangement is shown in figure 17-1. For safety considerations, explosive materials are separated into these primary and secondary high explosive categories and weapons are designed to maximize the safety of personnel until point of delivery.

The use of a three-component arrangement in which the primary high explosive, detonator, or primer is physically separated from the detonation sequence until just after weapon launch or before weapon impact maximizes this safety perspective. As discussed in chapter 16, this is often accomplished by forcing the explosive train to pass through a small physical port that connects the detonator to the booster, called a safing and arming device. This port can be blocked until the device is actually used.

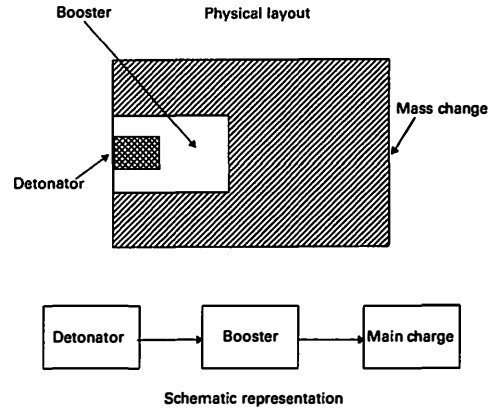


FIGURE 17-1 *High-explosive train sequence*

For example, this port might consist of two rotating plates with off-center holes. When the plates are aligned, the two holes line up and permit the detonation shock wave to transmit into the booster. This is called arming the device. Otherwise, the holes will not be aligned and even if the detonator goes off, the booster will remain safe because the shock wave will be prevented from being experienced in the booster itself. Other configurations exist, but all accomplish the same function of preventing inadvertent detonation and permitting detonation when desired.

As the first element in the explosive train, the detonator or primer is composed of a primary high explosive whose detonation is initiated by the fuze mechanism in the warhead. The shock wave produced from the detonator has both sufficient energy and proper alignment to transmit its energy into the booster charge.

Although classified as a secondary high explosive, the booster's sensitivity typically has an intermediary value between that of the primary high explosive in the detonator and the secondary high explosive of the main charge. As such, some texts create a third characterization of intermediate high explosives; this text, however, will not adhere to that. The booster adds its associated detonation wave energy to that of the detonator and directs it to actually initiate the main or bursting charge.

Although these three components form the basic high-explosive train, other components can be added as

part of the overall configuration, depending on what effect is desired. For example, a delay may be added between the detonator and the booster in order to prevent detonation of the booster for a specified period of time after the detonator has been fired.

Another typical component added to the high-explosive train is the relay, which may be required to strengthen the shock wave of the delay in order to initiate the booster. Table 17-1 shows the typical compounds used in the arrangement for the basic high-explosive train.

TABLE 17-1 *Common explosives and their uses*

Detonators/Primers	Booster	Main Bursting Charge
Mercury fulminate	Tetrytol	RDX
Lead azide	PETN	Comp—A, B, C
Lead styphnate	Tetryl	Cyclotol
Tetracene	TNT	HBX—1, 3
Diazodinitrophenol (DDNP)		H-6
		MINOL 2
		Ammonium picrate

Characteristics of Military Explosives

The characteristics of explosives used in military applications are considerably more stringent than those of explosives used in civilian applications. Whereas explosives used in construction, mining, or other engineering disciplines are used under tightly controlled, optimal conditions for the application of the explosive, military explosives are subjected to extreme ranges of operating conditions, including large temperature variations, pressure, and weather conditions. Consequently, through the years, many explosives have been examined for suitability for military use.

Although several have been found wanting, others have proved to be within acceptable standards. The six explosives presented in this text have been employed for many years, and are frequently used either by themselves or as core components of explosive mixtures in most of the chemical warheads in today's weapon inventory. The characteristics listed below are evaluated for each individual explosive material to determine its suitability for military applications.

Availability and cost Enormous quantities of explosives must be produced to meet military demands, and being able to derive these explosives from nonstrategic, inexpensive raw materials available in great quantities is highly desirable. Additionally, the manufacturing operations that gather, process, and deliver the final product must be reasonably simple, inexpensive, and safe.

Sensitivity This refers to the ease with which an explosive can be set off with one of the four basic methods of initiation: heat, pressure, shock, or electrostatic charge. The relative sensitivity of a given explosive to impact (shock) may vary greatly from its sensitivity to friction (heat); thus, each explosive is evaluated in a series of tests to determine its sensitivity to these various initiation methods.

These multiple empirical tests attempt to duplicate or mimic the accident environment that might be experienced during the handling of explosives. Furthermore, sensitivity is an important consideration in selecting the proper explosive to achieve a particular effect. Some of the testing methods used to determine sensitivity follow:

1. *Impact testing:* Sensitivity is expressed in terms of the distance through which a standard weight must be dropped to cause the material to react. In the drop-weight test, a small, fixed amount of explosive is subjected to a weight dropped upon it from various heights. The height at which a noticeable reaction is produced is recorded and tabulated against other explosive materials.

   Several different standard machines are used in the explosives industry, and criteria used in the tests vary as to the size, shape, weight dropped, total amount of sample used, and percentage of samples needed to produce the given "noticeable reaction" to conclude a success. Four of the most common machines used include the Picatinny Arsenal Machine (PAM), the Bureau of Mines Machine (BMM), the Los Alamos National Laboratory Machine (LANL), and the Lawrence Livermore National Laboratory Machine (LLNL). Table 17-2 lists composite results for the six common explosives presented in this text.

TABLE 17-2 *Impact sensitivity for common explosives*

Explosive	PAM (in.)	BMM (cm)	LANL/LLNL (m)
HMX	9	32	0.32
Nitroglycerin	7	15	0.20
PETN	6	17	0.13–0.16
RDX	8	32	0.28
Tetryl	8	26	0.37
TNT	14	95	1.48

2. *Friction:* Sensitivity is expressed in terms of the reaction produced when a weighted pendulum scrapes across the material. The friction pendulum test uses a pendulum with a 20-kilogram shoe of steel or fiber attached. The shoe is allowed to graze over a grooved anvil with a set sample (approx 7 grams) of explosive material, and the sample response is noted and recorded as either "explodes," "snaps," "crackles," or is "unaffected." Table 17-3 lists friction pendulum test results for our six explosives.

TABLE 17-3 *Friction pendulum test results*

Explosive	Steel Shoe	Fiber Shoe
HMX	Explodes	Unaffected
Nitroglycerin	Explodes	Explodes
PETN	Crackles	Unaffected
RDX	Explodes	Unaffected
Tetryl	Crackles	Unaffected
TNT	Unaffected	Unaffected

3. *Heat:* Sensitivity is expressed in terms of the temperature at which flashing or explosion of the material occurs. The explosion temperature test measures the exposure within a set interval that is necessary to set off a given explosive at various temperatures. Results are tabulated in table 17-4.
4. *Electrostatic Charge:* No widespread standard test is currently accepted for electrostatic spark testing of explosive sensitivity, although numerous tests have been devised and proposed.

Several other tests are frequently used to categorize explosive sensitivity, such as the rifle bullet impact test, the skid test, and the gap test.

TABLE 17-4 *Temperature (°C) for explosion test*

Explosive	0.1 sec	1.0 sec	5 sec	10 sec
HMX	380	—	327	306
Nitroglycerin	—	—	222	—
PETN	272	244	225	211
RDX	405	316	260	240
Tetryl	340	314	257	238
TNT	570	520	475	465

Stability Stability refers to the amount of time an explosive can remain in storage without undergoing serious deterioration or chemical decomposition. Stability is affected by the following three factors:

1. *Chemical composition:* Some common chemical compounds undergo deterioration and decomposition with time due to the inherent instability of their molecular structures. Certain organic molecular groups such as nitro dioxide (NO_2), nitrate (NO_3), and azide (N_3) intrinsically experience internal strain, and additional energy added in the form of heat or shock may cause a sudden disruption of the molecule and consequent explosion. In most cases, however, the internal strain of molecular instability simply results in degradation and decomposition of the explosive material over time at ordinary storage temperatures.
2. *Storage temperature:* Rates of explosive decomposition increase at higher temperatures, even for molecular arrangements not under internal strain, which is why shipboard magazines are temperature-controlled environments. All standard military explosives may be considered to be of a high order of stability at temperatures between −10° and +35°C, but there is a high-end temperature at which the rate of decomposition rapidly accelerates, thereby reducing stability. As a rule of thumb, most explosives become dangerously unstable at temperatures exceeding 70°C.
3. *Exposure to the sun's ultraviolet radiation:* The high-energy ultraviolet rays of the sun may result in rapid decomposition of many explosive compounds, particularly those containing nitrogen groups, discussed above. Generally, however, most explosives are sealed within weapon housings or casings, and therefore present a serious hazard only when a weapon has been damaged, resulting in exposure of the internal explosive to direct sunlight.

Power Power, or more properly, performance, refers to an explosive's ability to do work. In practice it is defined as an explosive's ability to accomplish what is intended in the way of energy delivery—that is, fragments, air blast, high-velocity jets, underwater bubble energy, and so on. Explosive power or performance is evaluated by a tailored series of tests to assess the material for its intended use. Of the tests listed below, cylinder expansion and air-blast

tests are common to most testing programs whereas the others support specific uses.

1. *Cylinder expansion test*: A standard amount of explosive is loaded into a cylinder usually made of copper. Data are collected concerning the rate of radial expansion of the cylinder and the maximum cylinder wall velocity. These data also establish the Gurney constant or ($\sqrt{2\Delta E}$ discussed in the next chapter).
2. *Cylinder fragmentation test*: A standard steel cylinder is charged with explosive and fired into a sawdust pit. The fragments are collected and the size distribution analyzed.
3. *Detonation pressure* (*Chapman-Jouget*): Detonation pressure data are derived from measurements of shock waves transmitted into water by the detonation of cylindrical explosive charges of a standard size.
4. *Determination of critical diameter*: This test establishes an explosive's minimum physical diametric dimension referred to as the critical diameter necessary to sustain its own detonation wave. Explosive charge diameters less than this minimum will deflagrage rather than detonate. The test procedure involves the detonation of a series of charges of different diameters until difficulty in detonation wave propagation is observed.
5. *Infinite diameter detonation velocity*: The speed at which the shock wave propagates through the explosive material, referred to as the detonation velocity, is dependent on density, charge diameter, and grain size. This procedure requires a series of charges of the same density and physical structure, but different diameters, to be fired and the resulting detonation velocities interpolated to predict the detonation velocity of a charge of infinite diameter.
6. *Pressure versus scaled distance*: A charge of specific size is detonated and its pressure effects measured at a standard distance. The values obtained are compared with that for TNT and can be predicted using the scaling law for conventional explosives (discussed in the next chapter).
7. *Impulse versus scaled distance*: A charge of specific size is detonated and its impulse—the area under the pressure-time curve—is measured versus distance. The results are tabulated and expressed in TNT equivalents.
8. *Relative bubble energy* (*RBE*): A 5–50 kg charge is detonated in water and piezoelectric gauges are used to measure peak pressure, time constant, impulse, and energy.

Brisance An explosive's ability to shatter a given material is referred to as its brisance, from the French, meaning "to break." This characteristic is of practical importance in evaluating an explosive's ability to shatter encased vessels such as bomb casings and grenades. In general, the rapidity with which an explosive reaches its peak pressure may be taken as a measure of its brisance; several different tests have been devised to evaluate this characteristic.

In the sand test, 0.4 gram of the explosive is buried in 200 grams of special sand and then detonated. The amount of sand shattered is used as an indication of the explosive's brisance, and provides a relative scale for measurement against other explosives evaluated in the same way. Other evaluation tests include the plate dent test, in which the size of a dent produced by the explosion is indicative of its brisance, or the explosive projectile test, where the number, size, and fragment velocities of an exploded projectile are used to measure brisance.

Density Density of loading refers to the unit weight of an explosive per unit volume, generally measured in grams per cubic centimeter. Several methods of loading are available, and the one used is determined by the characteristics of the explosive. The methods available include pellet loading, cast loading, or press loading.

Depending upon the method employed, an average density of the loaded charge can be obtained that is within 80%–95% of the theoretical maximum density of the explosive. High load density can reduce explosive sensitivity by making the mass more resistant to internal friction. If density is increased to the extent that individual crystals are crushed, the explosive will become more sensitive. Increased load density also permits the use of more explosive, thereby increasing warhead strength.

Volatility Volatility, or the readiness with which a substance vaporizes, is an undesirable characteristic of military explosives since excessive volatility often results in the development of pressure within rounds of ammunition and the separation of mixtures into their constituent components. Explosives must be no more than slightly volatile at the temperature at which they are loaded or at their highest storage temperature.

Stability, as mentioned before, is the ability of an explosive to withstand storage conditions without significant deterioration. Volatility affects the chemical composition of the explosive in such a way that a marked reduction in stability may occur, resulting in increased danger when handling the weapon.

Hygroscopicity Hygroscopicity is a measurement of a material's moisture absorbing tendencies. The introduction of moisture into an explosive is highly undesirable since it results in reduced sensitivity, strength, and detonation velocity of the explosive through the introduction of inert material that reduces the continuity of the explosive mass.

Moisture also affects explosives adversely by absorbing heat during detonation; by acting as a solvent medium for the explosive, causing chemical decomposition; and by promoting undesirable chemical side reactions with corrosive interaction of the explosive's metal container. Although the moisture content evaporates upon detonation, the energy loss due to absorption of the latent heat of vaporization of liquid water results in a reduced temperature of reaction and an overall decrease in the explosive's efficiency. For all of these reasons, hygroscopicity must be negligible in military explosives.

Toxicity Most explosives are toxic to some extent due to their chemical structure. Since toxic effects of various explosive materials vary widely from mild headaches to serious damage of internal organs, proper care must be taken to limit toxicity in military explosives to a minimum, including proper handling and wearing protective gear.

In most military applications, the explosive is sealed within a delivery container—for example, a shell or bomb casing—and will not present a serious hazard unless the weapon is physically damaged and its payload opened. Any explosive of high toxicity is unacceptable for military use.

Measurements of Chemical Explosive Reactions

The manufacture of new and improved explosives requires continuous research and development. Adoption of an explosive for a particular use is based upon both proving ground and service tests. Preliminary estimates of explosive characteristics are made using thermochemical principles.

Thermochemistry is a branch of chemistry concerned with internal changes in energy, principally heat, in chemical reactions. An explosion consists of a series of reactions, highly exothermic overall, that involve decomposition and recombination of the initial reactants to form the products of the explosion.

The total change in energy of the explosive reaction is calculated from known chemical laws and by-product analysis. For most reactions, tabular data from previous investigations permit rapid calculation of these energy changes. Since only the final products may be analyzed conveniently, indirect or theoretical methods are often used to determine the maximum temperature and pressure values.

Some important explosive characteristics that can be determined by such theoretical computations are:

1. Volume of explosion products (V)
2. Oxygen balance (OB%)
3. Heat of explosion ΔH^0_{exp}
4. Power index of the explosive (PI)

Regardless of the computation, balancing the chemical explosive reaction equation is required to fully evaluate these theoretical computations.

Explosive Reactions

Many chemical reactions release energy. If energy is liberated during the chemical reaction, the reaction is said to be exothermic. If the reaction proceeds slowly, the released energy will dissipate into the surrounding environment with few noticeable effects other than perhaps an increase in temperature. On the other hand, if the reaction proceeds very rapidly, a great quantity of energy can be released into a relatively small volume, manifesting itself as a rapid expansion of hot gases and creating a shock wave or propelling casing fragments outward at high speed.

Chemical explosions are distinguished from other exothermic chemical reactions by the extreme rapidity with which their reaction takes place. In addition to this violent release of energy, chemical explosions provide a means to transfer energy into mechanical work. This transference is accomplished through the rapidly expanding gaseous products created by the explosion, which forms a mechanical shock wave with sufficient force to damage nearby objects or shatter warhead casings.

The chemical explosion converts the solid or liquid explosive compound into a volume of gases 10,000 to 15,000 times greater than the original volume of the explosive. The resultant shock wave can reach velocities approaching 8,000 meters per second with associated temperatures of 3,000°C to 4,000°C. If no gaseous by-products are formed, then the liberated energy remains in either the solid or liquid products as heat, with no effective transfer of energy to the target.

Most chemical explosions are complicated, but they can be modeled as if they exhibit a set of straightforward reactions. To balance chemical explosive equations, we can apply the partial reactions shown in table 17-5, assuming that each partial reaction is completed to the maximum extent possible in the order of precedence listed in the table. For steps that do not relate to a given chemical

TABLE 17-5 *Order of priorities for explosive product decomposition reactions*

Priority	Explosive Composition		Decomposition Products
1	Metal + chlorine	→	Metallic chloride (solid)
2	Hydrogen + chlorine	→	Hydrogen chloride (HCl, gaseous)
3	Metal + oxygen	→	Metallic oxide (solid)
4	Carbon + oxygen	→	Carbon monoxide (CO, gaseous)
5	Hydrogen + oxygen	→	Water (H_20, gaseous)
6	Carbon monoxide + oxygen	→	Carbon dioxide (CO_2, gaseous)
7	Nitrogen, oxygen, hydrogen (excess)	→	N_2, O_2, H_2 (gaseous elemental state)

compound, we omit the application of that step and proceed to the next step in order.

The overall summation of the reaction is then totaled by adding all steps together. Since explosive reactions take place with great rapidity, only that oxygen present in the molecular structure of the explosive material—not atmospheric oxygen—is considered to participate in the explosion.

The following two examples are presented to demonstrate the application of table 17-5 to two common chemical explosives: TNT and nitroglycerin.

EXAMPLE 17-1: Balance the explosive combustion of TNT ($C_7H_5N_3O_6$).

SOLUTION: Priorities 1, 2, and 3 do not apply, because TNT contains no metal or chlorine.

Priority 4: Carbon + oxygen yields carbon monoxide (gaseous):

$$7C + 6O \rightarrow 6CO_{(g)}, \text{ leaving } 1C, 5H, 3N$$

Priority 5: Hydrogen + oxygen yields water (gaseous), which is not applicable because no oxygen remains (all oxygen was consumed during priority 4).

Priority 6: Carbon monoxide and oxygen yields carbon dioxide (gaseous), which is not applicable because no oxygen remains (all oxygen was consumed during priority 4).

Priority 7: Nitrogen (excess) yields N_2 (gaseous elemental state):

$$3N \rightarrow 1.5N_{2(g)}, \text{ leaving } 1C, 5H$$

Hydrogen (excess) yields H_2 (gaseous elemental state):

$$5H \rightarrow 2.5N_{2(g)}, \text{ leaving } 1C \text{ (solid elemental state)}$$

Overall reaction:

$$C_7H_5N_3O_{6(s)} \rightarrow 6CO_{(g)} + 1.5N_{2(g)} + 2.5H_{2(g)} + 1C_{(s)}$$

EXAMPLE 17-2: Balance the explosive combustion of nitroglycerin ($C_3H_5N_3O_9$).

SOLUTION: Priorities 1, 2, and 3 do not apply, because nitroglycerin contains no metal or chlorine.

Priority 4: Carbon + oxygen yields carbon monoxide (gaseous):

$$3C + 3O \rightarrow 3CO_{(g)}, \text{ leaving } 6O, 5H, 3N$$

Priority 5: Hydrogen + oxygen yields water (gaseous):

$$5H + 2.5O \rightarrow 2.5H_2O_{(g)}, \text{ leaving } 3.5O, 3N$$

Priority 6: Carbon monoxide and oxygen yields carbon dioxide (gaseous):

$$3CO_{(g)} + 3O \rightarrow 3CO_{2(g)}, \text{ leaving } 0.5O, 3N$$

Priority 7: Nitrogen (excess) yields N_2 (gaseous elemental state):

$$3N \rightarrow 1.5N_{2(g)}, \text{ leaving } 0.5O$$

$$0.5O \rightarrow 0.25O_{2(g)}$$

Overall reaction:

$$C_3H_5N_3O_{9\,(s)} \rightarrow 2.5H_2O_{(g)} + 3CO_{2\,(g)} + 1.5N_{2(g)} + 0.25O_{2\,(g)}$$

Although the six chemical explosives dealt with in this text contain neither metallic compounds nor chlorine, one may find that a given weapon in the U.S. arsenal indeed uses these materials in its explosive mixtures; hence, the overall table of priorities as presented in table 17-5 is left unaltered for this text.

Volume of Products of Explosion

Avogadro's law states that the volume of a gas maintained at constant temperature and pressure is directly proportional to the number of moles of the gas. From this law, it follows that the molar volume of one gas is equal to the molar volume of any other gas under identical conditions. The molar volume of any gas at 0°C and normal atmospheric pressure is very nearly 22.4 liters.

Thus, considering the balanced nitroglycerin reaction previously shown:

$$C_3H_5N_3O_{9(s)} \rightarrow 2.5H_2O_{(g)} + 3CO_{2(g)} + 1.5N_{2(g)} + 0.25O_{2(g)} \qquad (17\text{-}1)$$

The explosion of 1 mole of nitroglycerin produces in the gaseous state 2.5 moles of H_2O, 3 moles of CO_2, 1.5 moles of N_2, and 0.25 mole of O_2.

The decomposition of 1 mole of nitroglycerin therefore produces a total of 7.25 (3 + 2.5 + 1.5 + 0.25) total moles of gas. Assuming that these molecular volumes of gas are at standard temperature and pressure (STP, defined as 0°C and 1 atm pressure), then these gases occupy an actual volume of:

$$V_{STP} = 7.25 \text{ mol} \times \frac{22.4 \text{ liters}}{\text{mol}} = 162.4 \text{ liters} \qquad (17\text{-}2)$$

The density of nitroglycerin is 1.591 g/ml, so 1 mole would occupy a volume of 0.142 liter prior to the reaction as shown in equation 17-3:

$$\frac{227\text{g}}{\text{mol}} \times \frac{1\text{ml}}{1.591\text{g}} \times \frac{1\text{liter}}{1{,}000\text{ml}} = 0.142 \frac{\text{liter}}{\text{mol}} \qquad (17\text{-}3)$$

Upon completion of the reaction, we find an increase in volume of 1,160 times just at 0°C.

Based upon this simple beginning, the volume of the products of explosion can be predicted for any quantity of a given explosive at any temperature by applying Charles's law. Charles's law for perfect gases relates the total volume of gaseous products for any given temperature:

$$V_T = \left(\frac{V_{STP}}{273°\text{K}}\right) \times T \qquad (17\text{-}4)$$

where:

V_T = volume at temperature T
V_{STP} = Volume at STP
T = temperature in°K

The above formula is based on a perfect gas expanding 1/273 of its volume for each degree (1°K) rise in temperature above 0°C at a constant pressure. For example, if the above reaction for nitroglycerin were to reach 3,000°K, the molar volume occupied by the products of detonation of 1 mole of nitroglycerin would then be:

$$V_{3{,}000°\text{K}} = \left(\frac{162.4 \text{ L}}{273°\text{K}}\right) \times 3{,}000°\text{K} = 1{,}784 \text{ L}$$

This result, 1,784 liters, is approximately 12,500 times the initial volume of nitroglycerin.

In the case of TNT, the solid carbon did not react; see example 17-1. Therefore, the number of moles of gas formed in the TNT case was 10 moles, which equals 224 liters of gas at STP.

Although the total molar volume of gas produced may be fixed based upon the initial moles of explosive present, the damaging effects produced by the sudden transformation of the explosive compound into these gases—and the resultant shock wave produced and augmented by the high exothermic temperatures generated during the explosion—cannot be overstated.

Oxygen Balance (OB%)

Oxygen balance is an expression used to indicate not only the degree to which an explosive can be oxidized—that is, reacts with oxygen—but also the degree of efficiency of a particular chemical explosive. If an explosive molecule contains exactly enough oxygen to completely convert all of its carbon to carbon dioxide, all of its hydrogen to water, and all of its metal to metal oxide with no excess oxygen left over, the molecule is said to have a zero oxygen balance and is considered to be completely efficient in the use of its chemical material.

The molecule has a positive oxygen balance if it contains more oxygen than is needed and a negative oxygen balance if insufficient oxygen is present for the aforementioned reactions. In either case, the overall explosive reaction is less than ideal. In the oxygen-rich case, the excess oxygen represents a loss of physical density of explosive material by the additional space that the excess oxygen occupied, whereas in the oxygen deficient case, incomplete oxidation has occurred with some material left unreacted.

An explosive's sensitivity, strength, and brisance are all somewhat dependent upon oxygen balance and tend to approach their maximums as oxygen balance approaches zero.

The oxygen balance (OB%) is calculated from the explosive compound's empirical formula as the percentage of oxygen required for complete conversion of carbon to

carbon dioxide, hydrogen to water, and metal to metallic oxide. The calculation for oxygen balance is as follows:

$$OB\% = \frac{-16\ \text{g/mol}}{\text{compound's molecular weight}} \times \left(2C + \frac{H}{2} + M - O\right) \times 100 \quad (17\text{-}5)$$

where:

C = number of atoms of carbon
H = number of atoms of hydrogen
O = number of atoms of oxygen
M = number of atoms of metal

This calculation determines the number of grams of oxygen atoms that are in excess or deficient for 100 grams of an explosive compound.

Continuing with TNT ($C_7H_5N_3O_6$) and nitroglycerin ($C_3H_5N_3O_9$) as examples, the following calculations are provided:

EXAMPLE 17-3: Find the OB% for TNT.

SOLUTION: TNT molecular weight = 227.1 g/mol

C = 7
H = 5
O = 6
M = 0

$$OB\% = -\frac{1{,}600}{227.1}\left(2\times 7 + \frac{5}{2} + 0 - 6\right)$$
$$= -74\% \text{ for TNT}$$

EXAMPLE 17-4: Find the OB% for nitroglycerin.

SOLUTION: nitroglycerin molecular weight = 227.1 g/mol

C = 3
H = 5
O = 9
M = 0

$$OB\% = -\frac{1{,}600}{227.1}\left(2\times 3 + \frac{5}{2} + 0 - 9\right)$$
$$= +3.52\% \text{ for nitroglycerin}$$

Because sensitivity, brisance, and strength are properties resulting from a complex explosive chemical reaction, a simple relationship such as oxygen balance cannot be depended upon to yield universally consistent results. When using oxygen balance to predict properties of one explosive relative to another, it is to be expected that the one with an oxygen balance closer to zero will be the more brisant, powerful, and sensitive.

However, many exceptions to this rule do exist. More complicated predictive calculations, such as those discussed in the next section, result in more accurate predictions. One area in which oxygen balance can be applied is in the processing of explosive mixtures or composites.

Amatols are a family of explosives that are mixtures of ammonium nitrate and TNT. Ammonium nitrate has an oxygen balance of +20% whereas TNT, as calculated previously, has an oxygen balance of −74%. By creating an explosive mixture using one component that is oxygen rich and another that is oxygen deficient, a hybrid is created that is more powerful, sensitive, and brisant than either component individually.

In actual practice, a mixture of 80% ammonium nitrate and 20% TNT by weight yields an oxygen balance of +1%, the best properties of both mixtures, and an increase in power of 30% over TNT individually.

ANFO, which is a 94/6 mixture of ammonium nitrate and fuel oil, has a −0.6% oxygen balance. Composite explosives generally are designed to have oxygen balances that are closer to the ideal of zero, and many contain substances that are inert but provide stability for the mixture such as plasticizers or wax. See table 17-6 for examples of composite explosives.

Heat of Explosion

When a chemical compound is formed from its constituent components, the reaction may either absorb or liberate heat. The quantity of heat absorbed or given off during transformation is called the heat of formation. The heat of formation at the standard state is designated by the symbol ΔH_f^0. Heats of formation for solids and gases

TABLE 17-6 *Composite explosives*

Name	Composition	Formula
Amatol	80/20 ammonium nitrate/TNT	$C_{0.62}H_{4.44}N_{2.26}O_{3.53}$
ANFO	94/6 ammonium nitrate/#2 diesel oil	$C_{0.365}H_{4.713}N_{2.000}O_{3.000}$
Comp A-3	91/9 RDX/WAX	$C_{1.87}H_{3.74}N_{2.46}O_{2.46}$
Comp B-3	64/36 RDX/TNT	$C_{6.851}H_{8.750}N_{7.650}O_{9.300}$
Comp C-4	91/5.3/2.1/1.6 RDX/Di(2-ethyhexyl) sebacate/polyisobutylene/motor oil	$C_{1.82}H_{3.54}N_{2.48}O_{2.51}$

TABLE 17-7 *Heats of formation of common explosives and decomposition products*

Name	Molecular Formula	Molecular Weight (g/mol)	ΔH_f° (kJ/mol)
Aluminum chloride	$AlCl_3$	133.3	−696.4
Ammonium perchlorate	NH_4ClO_4	117.5	−289.0
Ammonium picrate	$C_6H_2(NO_2)_3O \bullet NH_4$	246.1	−326.8
Ammonium nitrate	NH_4NO_3	80.0	−367.9
Carbon monoxide	CO	28.0	−111.8
Carbon dioxide	CO_2	44.0	−393.5
Copper picrate	$[C_6H_2(NO_2)_3O]_2Cu$	44.0	−253.1
Cupic oxide	CuO	79.5	−146.2
RDX	$C_3H_6N_6O_6$	222.1	+83.8
HMX	$C_4H_8N_8O_8$	296.0	+104.8
Hydrochloric acid	HCl	36.5	−92.3
Lead azide	PbN_6	291.3	+443.7
Lead oxide	PbO	223.2	−219.8
Lead nitrate	$Pb(NO_3)_2$	331.2	−453.8
Lead styphnate	$C_6H(NO_2)_3(O_2Pb)$	450.3	−187.9
Mercuric oxide	HgO	216.6	−90.9
Mercury fulminate	$Hg(CNO)_2$	284.6	+270.3
Mercury picrate	$[C_6H_2(NO_2)_3O]_2Hg$	656.3	179.3
Metriol trinitrate	$C_5H_9N_3O_6$	255.1	−434.0
Nitroglycerin	$C_3H_5N_3O_9$	227.1	−333.7
Nitromethane	CH_3NO_2	61.0	−115.6
PETN	$C_5H_8N_4O_{12}$	316.2	−514.6
Picric acid	$C_6H_2(NO_2)_3OH$	229.1	−234.6
Tetryl	$C_7H_5N_5O_8$	287.1	+38.9
TNT	$C_7H_5N_3O_6$	227.1	−54.4
Water	H_2O	18.0	−240.6
Zinc oxide	ZnO	81.4	−353.4
Zinc picrate	$[C_6H_2(NO_2)_3O]_2Zn$	521.6	−430.7

found in explosive reactions have been determined for standard temperatures and atmospheric pressures and are presented in table 17-7; others may be found in a standard *CRC Handbook of Chemistry and Physics.*

Where a positive value is given, it indicates that heat is absorbed during the formation of the compound from its elements. Such a reaction is called an endothermic reaction.

Thermochemical calculations define heats of formations of all elements in their standard states and all temperatures as zero. Standard state is defined as "the state at which the elements are found under natural, ambient conditions."

The heat of formation of a compound is the net difference between the heat content of the compound and that of its elemental constituents. Since the elemental constituents are taken as zero by convention, it follows that the heat content of a compound is equal to its heat of formation in such nonrigorous calculations.

This leads to the principle of initial and final states, which may be expressed as follows: "The net quantity of heat liberated or absorbed in any chemical modification of a system depends solely upon the initial and final states of the system, provided the transformation takes place at constant volume or at constant pressure. It is completely independent of the intermediate transformations and of the time required for the reactions."

From this it follows that the heat liberated in any transformation accomplished through successive reactions is the algebraic sum of the heats liberated or absorbed in the different reactions. The net amount of heat liberated during an explosion, therefore, is the sum of the heats of formation of the products of the explosion minus the heat(s) of formation of the reactant(s) (the original explosive material).

The net heat difference between the heats of formations of the products and reactant in a chemical reaction

is termed the heat of reaction; for oxidation, this heat of reaction may be termed heat of combustion.

In explosive technology, only materials that are exothermic are of interest, and, hence, heats of reaction are virtually all negative. Since reactions may occur under conditions of either constant pressure or constant volume, the heat of reaction can be expressed at constant pressure or at constant volume. It is this heat of reaction that may be properly expressed as heat of explosion:

$$\Delta H^0_{expl} = \sum \Delta H^0_f \text{ (products)} - \sum \Delta H^0_f \text{ (reactants)} \quad (17\text{-}6)$$

where:

ΔH^0_{expl} = heat of explosion

ΔH^0_f(products) = heat of formation of products

ΔH^0_f(reactants) = heat of formation of reactants

EXAMPLE 17-5: Calculate the heat of explosion for TNT.

SOLUTION: From earlier, the balanced explosive combustion equation for TNT is:

$$C_7H_5N_3O_{6(s)} \rightarrow 6CO_{(g)} + 1.5N_{2(g)} + 2.5H_{2(g)} + 1C_{(s)}$$

Applying equation 17-6 yields:

$$\Delta H^0_{expl} = [6\times(-111.8) + 1.5\times(0) + 2.5\times(0) + 1\times(0)] - [-54.39] = -616.41 \text{ kJ/mol TNT}$$

The negative value indicates the liberation of heat (exothermic process) during this decomposition, which is consistent with the standard sign conventions used in chemistry. Diatomic molecular gases are considered to be standard state, and hence N_2, H_2, and O_2 have heats of formation equal to zero. Carbon in its solid form is also considered to be in its natural state and hence has a heat of formation equal to zero.

Expressed on a per mass basis, the value of heat of explosion obtains the symbol Q and is derived by taking the heat of explosion and dividing by the molecular weight or:

$$Q = \Delta H^0_{expl} \times \frac{1 \text{ mol explosive}}{\text{MW explosive}} \quad (17\text{-}7)$$

For TNT the heat of explosion per gram of TNT is calculated as follows:

$$Q = \left(-616.4 \frac{\text{kJ}}{\text{mol TNT}}\right) \times \left(\frac{1000 \text{ J}}{1 \text{ kJ}}\right) \times \left(\frac{\text{mol TNT}}{227 \text{ g TNT}}\right) = -2{,}715 \text{ J/g TNT}$$

Therefore TNT releases 2,715 Joules per gram of explosive. Performing the same calculations for nitroglycerin shows that this explosive releases 6,380.6 Joules per gram.

Explosive Power and Power Index

Although the heat of explosion is a useful comparative for various explosives, it does not take into account the volume change involved in the explosion. Combining the values of Q and volume of gas formed V_{STP}, both calculated independently, a value for the explosive power is given in the following equation.

$$\text{explosive power} = Q \times V_{STP} \quad (17\text{-}8)$$

It is useful then to compare various explosives to a reference explosive to determine the relative power as a percent between explosives. The reference explosive for convenience is usually selected as TNT. The result of the comparison is a ratio in a per mass basis of the explosive powers, called the power index.[i]

$$\text{power index} = \frac{Q \times V_{STP}}{Q(\text{TNT}) \times V_{STP}(\text{TNT})} \times 100 \quad (17\text{-}9)$$

Continuing with the example of TNT and nitroglycerin:

$$\text{power index} = \frac{-6{,}380.6 \frac{\text{J}}{\text{g}} \times 162.4 \text{ liters}}{-2{,}715 \frac{\text{J}}{\text{g}} \times 224 \text{ liters}} \times 100 = 170\%$$

Therefore nitroglycerin is 170% more explosive than a comparative mass of TNT.

[i] Akhavan, 85

18 Weapon Damage Effects

Introduction

The basic function of any weapon is to deliver a destructive force to an enemy target.[i] Targets of today include military bases, factories, bridges, ships, submarines, tanks, missile launching sites, deeply buried fortifications, and troop concentrations. Since each type of target presents a different physical destruction problem, a variety of general and special-purpose warheads are required, within the bounds of cost and logistical availability, so that each target may be attacked with maximum effectiveness.

When a warhead detonates in the vicinity of a target, the expectation is that the target will be damaged to some extent. However, there is no guarantee that the target will be destroyed or incapacitated. There are too many factors involved that may alter the outcome of the engagement. Therefore, the prediction of the amount of damage to a target must be expressed as a probability.

Quantifying Damage

In order to make sense out of the infinite spectrum of outcomes, the use of black-and-white terms becomes useful. For most military engagements, it matters only that the target is removed from action. It is in this sense that the target is considered damaged. The probability of damage, P_d, is a statistical measure of the likelihood that the target will in some way be incapacitated and possibly destroyed. P_d will depend on the proximity of the warhead to the target, the vulnerability of the target, and the effects of the warhead—that is, shock wave, fragments, and so on.

The proximity of a warhead to the target is a measure of the accuracy of the delivery vehicle, usually described by the amount of circular error probable, (CEP). Target vulnerability and warhead effects go hand in hand because every target will be affected differently by different weapons.



FIGURE 18-1 *JDAM is an inertially guided bomb with GPS capability.*

In actual application there are many other factors contributing to the overall P_d. These factors include all of the many components of the fire control problem—for example, detection, localization, launching, guidance, fuzing capabilities, and warhead effectiveness. Each of these factors adds conditional terms to the overall P_d.

In weapon and strike planning, much time and effort are spent precalculating weapon effectiveness versus a particular type of target. These data are consolidated in a series of publications called the *Joint Munitions Effectiveness Manuals* or *JMEMs*. *JMEMs* consist of more than a hundred volumes of data and instructions that include probability of damage, target vulnerability, and weapon accuracy data. Thankfully most users need to use only a few of the volumes and much of these data have been integrated into computer software.

Damage Description

Targets will usually have different damage levels. Each type of damage has an associated probability of occurrence for a specific weapon and number of hits by that weapon.

TABLE 18-1: *P_d for an unguided 500-pound bomb versus a vehicle*

Light	Minor damage, some functions lost, but still capable of operation	0.9
Moderate	Extensive damage, many functions lost, operation still possible but at reduced effectiveness	0.5
Heavy	Unable to operate	0.1

In the example provided in table 18-1, light damage will result in some lost functionality and has a 90% probability of occurring given a hit. It should be clear that this does not mean that 90% of the target is lightly damaged.

Other categories of damage might describe the ability of an aircraft to take off, a missile to be launched, or a vehicle to be mobile.

A different target would have different criteria as seen in table 18-2.

TABLE 18-2: *P_d for an unguided 500-pound bomb versus a mobile missile launcher*

Damage Level	Description	P_d
Mobility	Vehicle unable to move	0.7
Firepower	Unable to launch missile	0.4
Catastrophic	Extensive damage	0.1

One way to improve on this situation would be to have a more accurate weapon. If a GPS-guided bomb was launched, a great improvement could be made, as seen in table 18-3.

TABLE 18-3: *P_d for a GPS-guided 500-pound bomb versus a mobile missile launcher*

Damage Level	Description	P_d
Mobility	Vehicle unable to move	0.9
Firepower	Unable to launch missile	0.9
Catastrophic	Extensive damage	0.9

Note that the probability of damage is the same for all three damage levels, so there is a 90% chance that the target will suffer damage. This is due to the greatly improved accuracy of the weapon. By improving accuracy, lethality is improved. To examine the probability of damage more closely, it is broken into its separate variables.

$$P_d = P_{d/h} \times P_h \tag{18-1}$$

where:

$P_{d/h}$ = probability of damage given a hit
P_h = probability of a hit

As mentioned earlier, there are two components that contribute to probability of damage. The first is the probability of damage given a hit, which is a function of target vulnerability and weapon lethality. The second is the probability of a hit, which is a function of accuracy or circular error probable.

Circular Error Probable

The proximity of the warhead to its intended target is also statistical in nature. What is of interest is the most likely distance from the point of impact and/or detonation to the target, if many warheads were launched at it. The most probable outcome from a single launch is a function of this value.

If the probability of impacting within a given distance of the target in any direction is precisely 50%, then that distance is referred to as the circular error probable. Alternatively, CEP can be defined as:

> CEP = the radius of a circle about the aim point inside of which there is a 50% chance that the weapon will impact and/or detonate.

The following discussion develops the assumptions and subsequent mathematical equation for CEP.

Discrete and Continuous Distributions

As individual rounds impact around an intended target, the miss distance from that target is said to be a discrete variable. A set of shots will be distributed over the face of the target, and discrete distribution will then describe how the shots are arranged.

Suppose five rounds are shot at a target and the miss distance is measured in Cartesian coordinates. See table 18-4.

There are two distributions: one for X and one for Y. Consider the X coordinate; the distribution could be made as shown in table 18-5.

Table 18-5 is known as a histogram when plotted. Note that the range of coordinates that make up the "bins"

TABLE 18-4: *Data from five rounds*

Shot Number	X Coordinate	Y Coordinate
1	2	3
2	3	1
3	−1	3
4	1	4
5	0	2

is arbitrary. In general, there should be enough bins to give representative distribution, but not so many that a lot of bins have 0s or 1s in them.

TABLE 18-5: *Distribution of five rounds*

Range of X Coordinates	Number of Shots in Specific Range
$\|X\| < 0$	1
$0 > \|X\| > 2$	3
$\|X\| > 2$	1

Discrete distributions have descriptive statistics associated with them:

1. The mean (average) describes the center point.
2. The standard deviation describes the variation about the mean.

The standard deviation is the measure of the average distance from the center without regard for the sign. One interpretation is that it also represents the expected variation in a single shot—that is, how far from center the shot is expected to hit.

This is exactly the quantity of interest for weapon designers because it represents the accuracy that can be expected. To compute the mean and standard deviation from a discrete distribution, we use the following well-known equations:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{18-2}$$

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2} \tag{18-3}$$

In the previous simple set of data:

$$\bar{x} = \frac{1}{5}(2+3+(-1)+1+0) = 1.0$$

and

$$\sigma_x = \sqrt{\frac{1}{5}((1-2)^2+(1-3)^2+(1-(-1))^2+(1-1)^2+(1-0)^2)}$$
$$= 1.41$$

In order to understand the behavior of distributions, it is useful to be able to describe them using analytic functions. This gives rise to the continuous distribution function, which is just a limiting case where there are a great number of sample points, thus allowing the use of smaller and smaller bins.

If the number of shots in a particular bin is represented as x_i, then the fundamental relationship between discrete and continuous distributions is:

$$x_i = f(x_i)dx \tag{18-4}$$

The distribution function $f(x)$ is a smooth approximation to the discrete case. It can be shown that many random events result in the same distribution function for $f(x)$, called the normal distribution and having the form:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}} \tag{18-5}$$

Figure 18-2 is a representation.

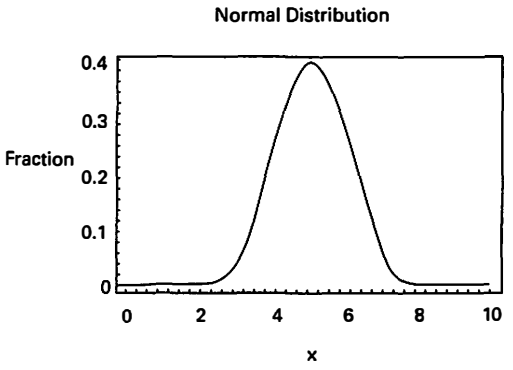


FIGURE 18-2 *Normal distribution*

This case is shown with average $x = 5$ and standard deviation $\sigma_x = 1$. The vertical axis displays the fraction of the total, where the distribution has been normalized so that the integral over all space = 1.

The CEP can be determined assuming that distribution of x and y are normal. If the x and y distributions are combined, they form a bivariate distribution. In the case where x and y have identical distributions, the circular

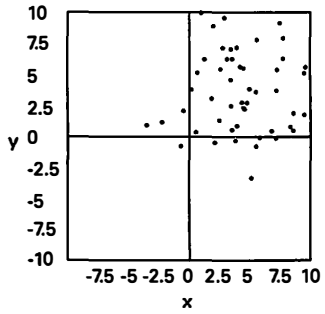


FIGURE 18-3 *Plot of miss distance*

error probable (CEP) can be computed from the combined bivariate distribution.

If they are close, but not identical, then they can be made identical by taking an average of the standard deviations, $\sigma = (\sigma_x + \sigma_y)/2$. Recall that the CEP is the radius of a circle about the aim point within which there is a 50% probability of hitting. In this case,

$$\text{CEP} = 1.17\sigma \tag{18-6}$$

Figure 18-3 is a plot of a weapon's miss distance. There are two types of errors experienced by this weapon. The first is random scattering about the aim point, called random error. The second is the concentration of rounds falling in the upper-right quadrant; this is known as systematic error.

From these data, the pertinent statistics are as follows:

$$\bar{x} = 4.2 \quad \sigma_x = 3.06$$
$$\bar{y} = 3.6 \quad \sigma_y = 3.20$$

Random Error

Since σ_x and σ_y are different, an average standard deviation is needed:

$$\sigma_R = \frac{3.06 + 3.20}{2} = 3.13 \tag{18-7}$$

This standard deviation represents the random (R) variation about the center of all the shots or mean impact point (MIP). This variation is due to a number of factors, one of which is the shooter's inability to fire the weapon when the sights are on the aim point. Another is the variability in the size and shape of each round.

The radius of a random error about the MIP is determined in the same manner as the CEP, or:

$$\text{radius about MIP} = 1.17\sigma_R \tag{18-8}$$

In the example given, 1.17(3.13) = 3.66 represents the radius of a circle centered about $(\bar{x},\bar{y})$ within which 50% of the shots lie. If random error were the only error, $\bar{x}$ and $\bar{y}$ would approach the aim point (0, 0) as the number of shots increases, and the CEP and the radius of the random error circle would be identical. Random error, however, is not the only error and the CEP must account for all of the errors associated with the weapon system.

Systematic Error

Since the aim in this example is the center and not to the right and up, there is a need to take into account this error, too. In weapon systems, this type of error can come only from a systematic source. "Systematic" implies that the exact error direction is unknown, but it will generally be in a specific amount and direction for each and every round. This comes from calibration error, windage, and many other sources. Although the error is consistent and can be corrected for, it must be included as an error since it is not known beforehand.

The displacement of the center of all shots from the aim point is a measure of the systematic error, which for our example is given by

$$\sigma_S = \sqrt{\bar{x}^2 + \bar{y}^2} = \sqrt{4.2^2 + 3.6^2} = 5.53$$

Now there are two standard deviations: one for the systematic error, σ_S, and one for the random error, σ_R. It can be shown that the two errors are independent and result in an overall distribution with a standard deviation of:

$$\sigma = \sqrt{\sigma_S^2 + \sigma_R^2} \tag{18-9}$$

And now CEP = 1.17σ as usual.

To further illustrate this, examine the sample data. Figure 18-4 shows the x and y data overlaid with the continuous distribution function found from the descriptive statistics. Note that the data have the general shape but still have many variations from the ideal case.

Using these data and the displacement of the center from the aim point, the systematic error, CEP = 7.44, results about the center. Figure 18-5 is a plot showing two circles:

1. The random variation about the center of the shots
2. The combined total error represented by the CEP

If you count the shots, you will find that approximately 50% of the shots are within the CEP.

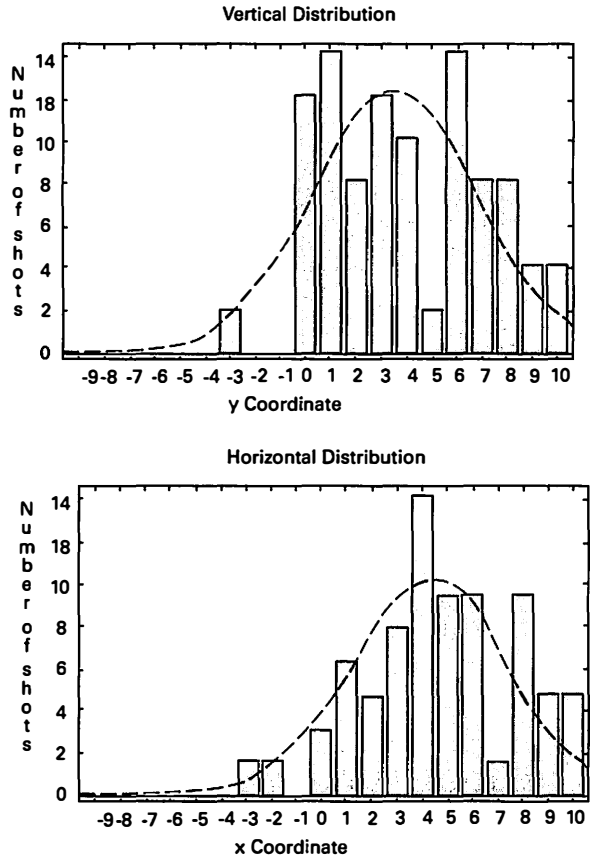


FIGURE 18-4 *Data for x and y*

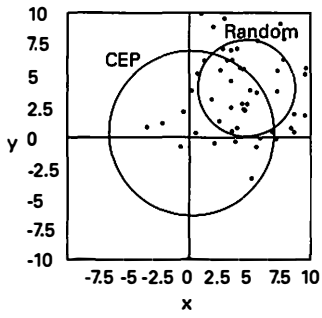


FIGURE 18-5 *CEP and random error circles*

Warhead Characteristics

The transfer of destructive energy from the warhead to the target is known as energy coupling. The energy forms liberated when a payload detonates include thermal, kinetic, chemical, and nuclear. In addition, pressure or shock waves are a by-product of the blast. The effectiveness of the energy transfer depends on many factors including target location and defenses; whether the energy source is chemical or nuclear; strength, power, and efficiency of the explosive agent; attenuation and propagation pathways; and the loss of efficiency due to environmental conditions. The effective damage to the target, then, is directly related to three parameters: damage volume, attenuation, and propagation.

Damage Volume

The warhead, as the focal point of the destructive agent, may be thought of as being enclosed in an envelope that sweeps along its trajectory. The volume enclosed by this envelope defines the limit of the destructive effectiveness of the payload.

Attenuation

As shock waves and fragments leave the point of origin, a reduction in their destructive potential per unit area takes place. Attenuation can be likened to an expanding sphere in which the energy available per unit area constantly decreases until it is completely harmless.

Propagation

This is the manner in which energy and material, emitted by the warhead at detonation, travel through the medium in which the blast occurs. When the propagation of a payload is uniform in all directions, it is called isotropic. If not, it is called nonisotropic; see figure 18-6.

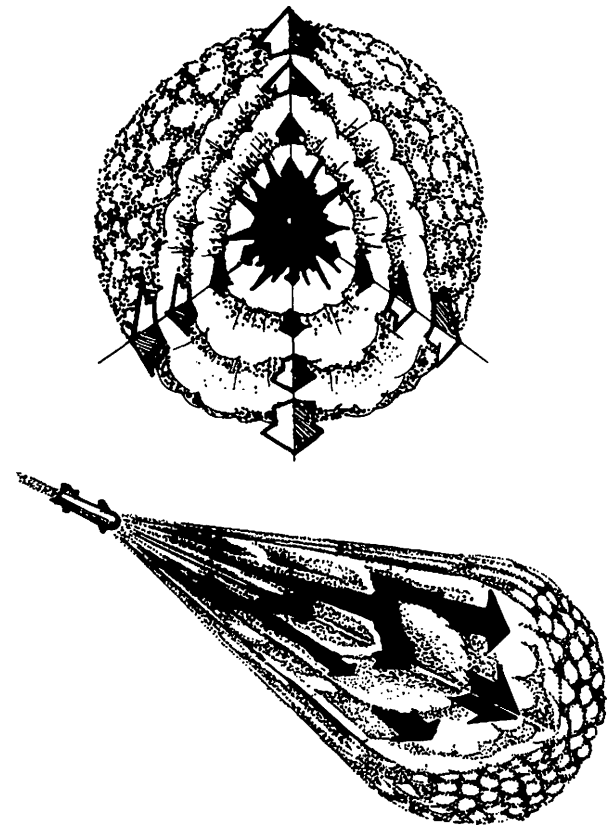


FIGURE 18-6 *Isotropic (top) and nonisotropic (bottom) blast patterns*

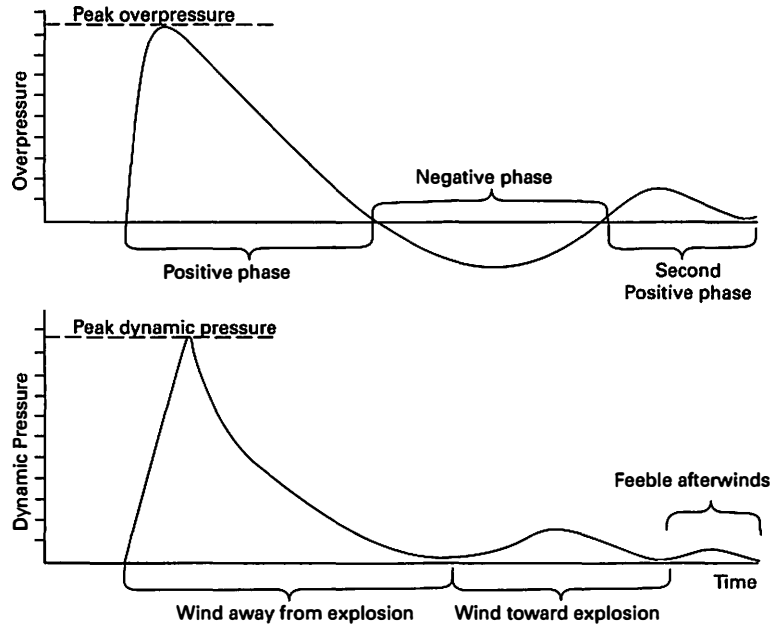


FIGURE 18-7 *Peak overpressure and dynamic pressure versus time relationship of a blast wave at a given distance from the blast center*

Blast Warheads

A blast warhead is one that is designed to achieve target damage primarily from blast effects. When a high explosive detonates, it is converted almost instantly into a gas at very high pressure and temperature. Under the pressure of the gases thus generated, the weapon case expands and breaks into fragments. The surrounding air is compressed and a shock (blast) wave is transmitted into it. Typical initial values for a high-explosive weapon are 200 kilobars of pressure (1 bar = 1 atmosphere) and 4,000°C.

The shock wave generated by the explosion is a compression wave in which pressure rises from atmospheric pressure to peak overpressure in a fraction of a microsecond. It is followed by a much slower (hundredths of a second) decline to atmospheric pressure. This portion is known as the positive phase of the shock wave. The pressure continues to decline to subatmospheric pressure and then returns to normal. This portion is called the negative or suction phase.

The durations of these two phases are referred to as the positive and negative durations. The area under the pressure-time curve during the positive phase represents the positive impulse and that during the negative phase represents the negative impulse. The result of this positive/negative pressure variation is a push/pull effect upon the target, which causes targets with large volumes to effectively explode from the internal pressure. Due to the pressure differential within the shock wave, the air will flow from high to low pressure. This creates a blast wind, which can be of substantial velocity, well over 100 mph. The blast wind lasts for only a fraction of a second and changes direction during the negative phase. A pressure-time curve is shown in figure 18-7 for both overpressure and dynamic pressure cases.

As the wind flows against objects, they feel dynamic pressure from the drag; see figure 18-8. The dynamic pressure experienced by an object follows the familiar equation for drag:

$$P_{dyn} = \frac{1}{2} C_d \rho v^2 \qquad (18\text{-}10)$$

where:

C_d = coefficient of drag for the particular object
ρ = density of air (normally ≈ 1.2 kg/m^3)
v = velocity of the blast wind

The peak overpressure is the instantaneous rise in pressure at the shock front and is the highest pressure in the shock wave. The positive phase is generally of short duration, less than 5/100th of a second. The negative phase lasts five to six times longer than the positive phase but is only a fraction of the maximum positive pressure.

As the shock wave moves outward from the center of the explosion, its peak overpressure decreases quickly. One reason for this decrease in energy is losses due to

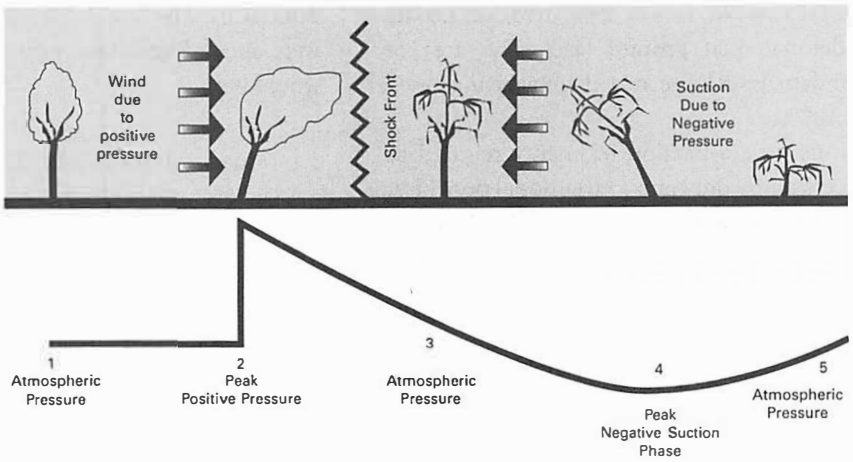


FIGURE 18-8 *Effects of a blast wave at a given distance from the blast center*

heating the surrounding air as the shock wave passes. Another is the rate of expansion of the volume of gases behind the blast wave. If the blast pressure expands spherically outward from its point of origin, it will decrease as the volume of the sphere increases.

In other words, the blast pressure is inversely proportional to the cube of the distance from the blast center ($1/R^3$). Under actual conditions, cooling effects, nonideal expansions, actual rupturing of the warhead casing, and subsequent fragment generation cause the attenuation factor to be greater than is indicated by theory.

Another aspect of overpressure occurring in air bursts is the phenomenon of Mach reflections, called the Mach effect. Figure 18-9 portrays an air burst at some unspecified distance above a reflecting surface, at five successive time intervals after detonation.

When a bomb is detonated at some distance above the ground, the reflected wave catches up to and combines with the original shock wave, called the incident wave, to form a third wave that has a nearly vertical front at ground level. This third wave is called a Mach wave or Mach stem, and the point at which the three waves intersect is called the triple point.

The Mach stem grows in height as it spreads laterally, and as the Mach stem grows, the triple point rises, described by a curve in the air. In the Mach stem the incident wave is reinforced by the reflected wave, and both peak overpressure and impulse are at a maximum that is considerably higher than the peak pressure and impulse of the original shock wave at the same distance from the point of explosion.

Using the phenomenon of Mach reflections, it is possible to increase considerably the radius of effectiveness of a bomb. By detonating a warhead at some optimum height of burst above the ground, the maximum radius at which a given pressure or impulse is exerted can be

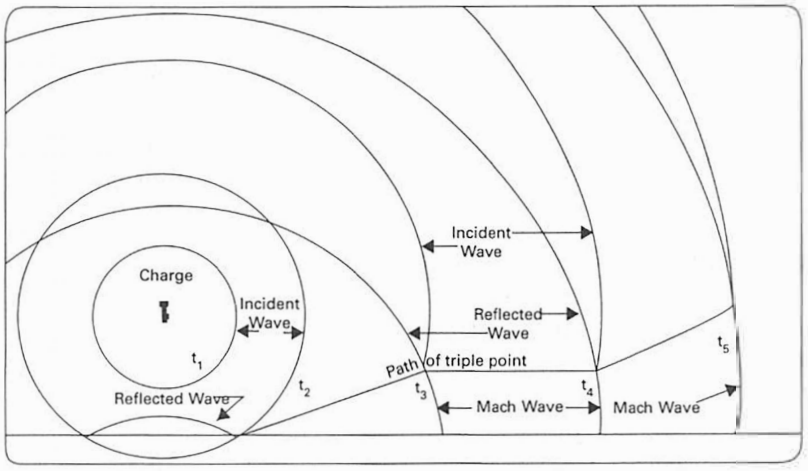


FIGURE 18-9 *Reflection of the blast wave from a surface and the resulting Mach stem*

increased, in some cases by almost 50%, over that for the same bomb detonated at ground level. The area of effectiveness, or damage volume, may thereby be increased by as much as 100%.

All explosions have a portion of energy released as blast, however, currently only one conventional pure blast warhead is in use, the fuel-air-explosive (FAE). Of course, all nuclear warheads are blast warheads, and on most targets they would be detonated at an altitude to make use of this Mach stem effect.

Blast effects damage targets in two primary ways called diffraction loading and drag loading. Diffraction loading is caused by the peak overpressure of the blast front rapidly applying pressure to a target from all sides including the roof as the shock wave passes. The shock wave diffracts around the structure, essentially squeezing it from multiple directions simultaneously; see figure 18-10.

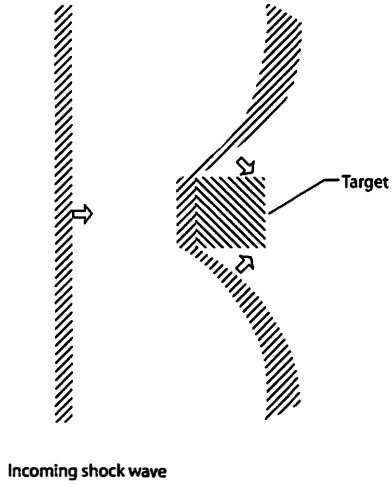


FIGURE 18-10 *Diffraction loading*

This diffraction results in a rapid buildup of pressure on the target to at least twice the overpressure of the initial shock front. Brittle material such as glass and concrete will shatter, whereas ductile material such as wood and metal will be crushed and bent from this pressure. The loading that a target experiences can be estimated from the peak overpressure, as shown in the following example.

EXAMPLE 18-1: Consider a shock wave of 25 psi peak overpressure incident upon a standard one-story residential house. The surface area of the house is estimated as follows:

Frontal area = 40 ft × 10 ft
Side area = 25 ft × 10 ft
Roof area = 40 ft × 13.3 ft (front half only of roofline at 20°)
Total area = 1,432 ft^2 × (144 $in.^2/ft^2$) = 206,200 $in.^2$

SOLUTION: The total loading, if all the peak overpressure were applied simultaneously to the front sides and roof, would be:

$$\text{load} = 25\,\text{psi} \times 206{,}200\,\text{in.}^2 = 5.2 \times 10^6\ \text{lbs}$$

This is a load of roughly 2,600 tons.

Drag loading on a target is caused by dynamic pressure. This type of structural loading is the aerodynamic force that acts on surfaces normal to the shock front. Because dynamic pressure is usually less than overpressure, the structural loading is reduced. However, because dynamic pressure is applied to the structure for a longer period of time and changes direction, the effect is still significant; see figure 18-11.

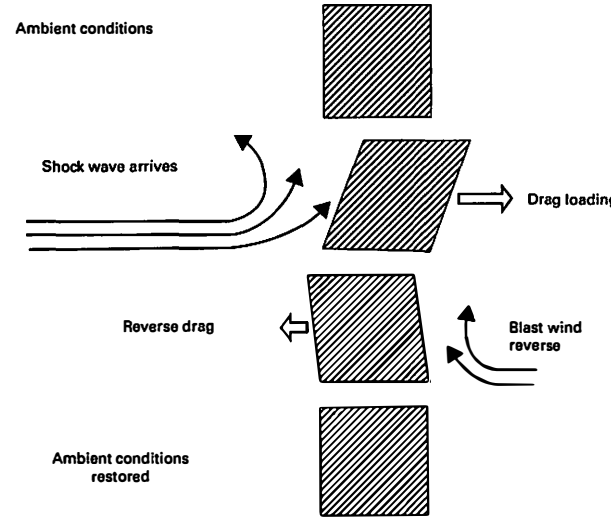


FIGURE 18-11 *Drag loading*

Some targets, those that are relatively flexible, are not damaged by diffraction loading. These same targets, however, may be vulnerable to drag loading damage. Targets that are not rigidly affixed will be thrown by the force and may be displaced by several meters. Personnel are very vulnerable to this type of damage as well as to the secondary threat of being hit by other objects and debris thrown by the blast. Aircraft and light equipment are also likely to be damaged by drag loading.[ii]

EXAMPLE 18-2: Calculate the drag loading using the same one-story residential structure as in example 18-1 and applying 3 psi of dynamic pressure.

SOLUTION: The structure would experience:

$$\begin{aligned}\text{drag load} &= \text{dynamic pressure} \times \text{frontal area} \\ &= (3\,\text{psi})(40\,\text{ft} \times 10\,\text{ft})(144\,\text{in.}^2/\text{ft}^2) \\ &= 172{,}800\,\text{lbs}\end{aligned}$$

This is a load of roughly 86 tons.

Due to the complex nature of explosions, it is not possible to easily predict the magnitude of these blast effects. However, there is a vast collection of experimental data from the explosion of 1 kilogram of TNT, which has been chosen as the reference explosion.

The values for an arbitrary explosion can be found by relating it to the reference explosion through a relation known as the scaling law. It relates the distances at which the same effect will be felt for different explosive amounts. The scaling factor is $W^{1/3}$, where W = the equivalent amount of TNT (in kilogram). W is found by multiplying the mass of the explosive by its power index (PI). Explicitly,

$$d_w = d_1 \times W^{1/3} \qquad (18\text{-}11)$$

where:

d_1 = distance from 1 kilogram TNT
d_w = distance from the W kilograms of TNT equivalent

Equation 18-11 holds true for both peak overpressure and dynamic pressure. The peak overpressure and dynamic pressure are found from the graph in figure 18-12 for 1 kilogram of TNT.

EXAMPLE 18-3: To understand how to use equation 18-11 and figure 18-12, consider the following example. Calculate the peak overpressure and dynamic pressure felt by a person facing a blast of 10 kilograms of TNT equivalent, standing 5 meters away.

SOLUTION: To use the graph, the distance must be scaled down to the reference condition.

$$d_1 = \frac{d_w}{W^{1/3}} = \frac{5\,\text{m}}{10^{1/3}} = 2.3\,\text{m}$$

using the graph,

peak overpressure = 23 psi
dynamic pressure = 2.0 psi

EXAMPLE 18-4: Find the distance from which light armor might be destroyed by a 2,000-pound TNT equivalent bomb.

SOLUTION: Referring to the dynamic pressure versus range curve for a reference explosion of 1 kilogram of TNT, 7 psi of dynamic pressure will be felt at about 1.5 meter.

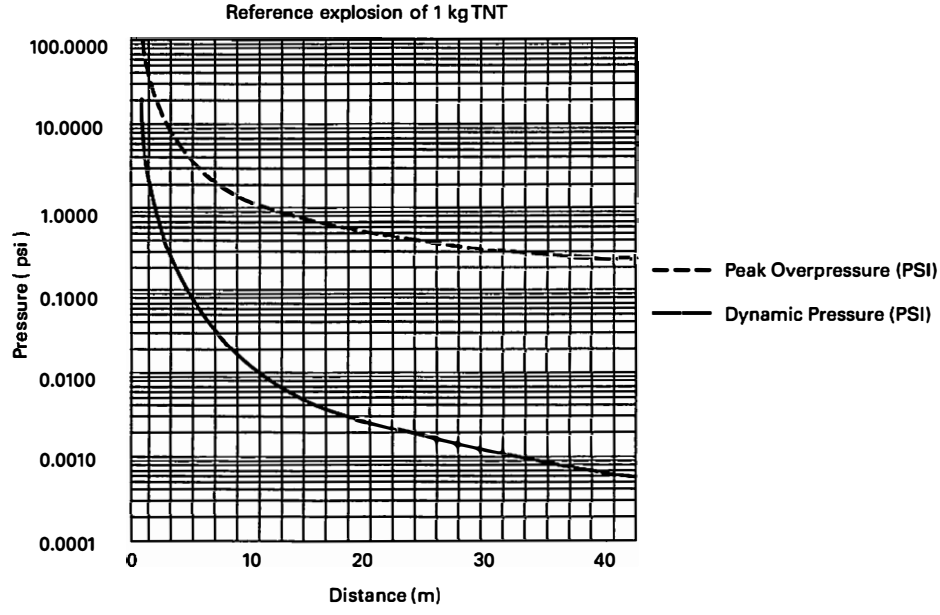


FIGURE 18-12 *Peak overpressure and dynamic pressure for 1 kilogram of TNT*

TABLE 18-6 *Sample damage criteria for blast effects*

Target	Damage Mechanism	Light (psi)	Moderate (psi)	Heavy (psi)
Industrial buildings	Diffraction	3	5	15
Roads and bridges	Diffraction	5	8	12
Light armor	Drag	1	4	7
Heavy armor	Diffraction	10	100	200
Troops in open field	Drag	1	3	5
Troops in bunkers	Diffraction	5	30	100
Shallow, buried structures	Diffraction	30	175	300
Parked aircraft	Drag	0.7	1.5	3
Ships	Drag	2	5	7

Table 18-6 is presented purely for academic purposes. This table lists possible targets and the various damage mechanisms and pressures needed to achieve an assumed level of damage. Actual tests and rigorous study in structures are required to determine accurate values on specific targets.

To find the scaling factor $W^{1/3}$, convert the warhead size into kilograms:

$$2{,}000 \text{ lb} = 910 \text{ kg TNT}$$
$$W^{1/3} = 910^{1/3} = 9.7$$

Therefore, the effect (7 psi dynamic pressure) will be felt to a range of:

$$d_{910} = d_1 \times W^{1/3} = 1.5 \text{ m} \times 9.7 = 14.5 \text{ m}$$

Fragmentation Warheads

When a charge of high explosive detonates within a closed container, the container is separated into many fragments. These fragments are hurled outward at high velocities and, in effect, become projectiles capable of inflicting damage on objects within their lethal range.

The study of ballistics, the science of the motion of projectiles, has contributed significantly to the design of fragmentation warheads. Specifically, terminal ballistics studies attempt to determine the laws and conditions governing the velocity and distribution of fragments, the sizes and shapes that result from bursting different containers, and the damage aspects of the bursting charge.

Fragmentation warheads are used principally as a means of destroying air targets, producing personnel casualties, and attacking light military equipment.

Approximately 30% of the energy released by the explosive detonation is used to fragment the case and impart kinetic energy to the fragments. The balance of available energy is used to create a shock front and blast effects. The fragments are propelled at high velocity, and after a short distance they overtake and pass through the shock wave.

The rate at which the velocity of the shock front accompanying the blast decreases is generally much greater than the decrease in the velocity of fragments, which occurs due to air friction. Therefore, the advance of the shock front lags behind that of the fragments. The radius of effective fragment damage, although target dependent, usually exceeds considerably the radius of effective blast damage in an air burst.

Whereas the effects of an idealized blast payload are attenuated by a factor roughly equal to $1/R^3$ (the radius R is measured from the origin), the attenuation of idealized fragmentation effects will vary as $1/R^2$ and $1/R$, depending upon the specific design of the payload.

Herein lies the principal advantage of a fragmentation payload: it can afford a greater miss distance and still remain effective because its attenuation is lower.

The disadvantage of this type of warhead is that it does not achieve the sudden, spectacular results of a blast warhead since the holes produced in a target by fragmentation may not have a crippling effect upon the target, particularly at the maximum range of fragment effect.

The velocity of the fragments can be looked at in two parts: first, initial fragment velocity, and second, velocity as a function of distance from the origin.

The initial velocity of the fragments of a warhead depends primarily upon three factors:

1. Charge-to-metal ratio, C/M, where C is the mass of explosive per unit length of the projectile and M is the mass of metal per unit length of the projectile
2. Characteristics of the explosive filler, particularly its brisance and strength
3. Shape of the fragmenting warhead

TABLE 18-7 *Gurney constants*

Explosive	Gurney Constant (m/s)
TNT	2,320
H-6	2,350
Composition B	2,680
Octol	2,560
PBX-9404	2,637

Expressing this quantitatively:

$$V_0 = \sqrt{2\Delta E}\sqrt{\frac{\frac{C}{M}}{1+K\left(\frac{C}{M}\right)}} \qquad (18\text{-}12)$$

where:

$\sqrt{2\Delta E}$ = the Gurney constant for the explosive material. The E term is the heat of explosion in J/kg. The Gurney constant has units of velocity, (m/s). (It is a rough measure of the speed of the explosion and is related to the potential energy of the given explosive; see table 18-7 for typical Gurney constants.)

C = explosive weight

M = weight of the casing

K = configuration constant (flat plate =1/3; cylinder = 1/2; sphere = 3/5)

TABLE 18-8 *Fragment velocity as a function of warhead wall thickness and C/M ratio for TNT*

Wall Thickness (cm)	Charge-to-Metal Ratio (C/M)	Initial Velocity V_0 (m/s)
1.27	0.165	875
0.95	0.231	988
0.79	0.286	1,158
0.48	0.500	1,859

Table 18-8 illustrates the relationship between the charge-to-metal ratio and the initial velocities (V_0) of the fragments. In this case, cylinders of 5.1 cm internal diameter, filled with TNT, were employed. Notice that as the charge-to-metal ratio increases, the fragment velocity also increases.

As fragments are flung outward from the casing, air resistance (drag) causes fragment velocity to start dropping. Solving the equations of motion for the projectile, the fragment's velocity as a function of distance (S) traveled is:

$$V_S = V_0 e^{-C_D \rho \frac{A}{2m} S} \qquad (18\text{-}13)$$

where:

C_d = coefficient of drag (depends on the shape of the fragment and, to some extent, the velocity)

ρ = density of air (normally 1.2 kg/m^3)

V_0 = fragment initial velocity (m/s)

A = cross-sectional area of the fragment (m^2)

m = mass of individual fragment (kg)

Thus, during its flight through the air, the velocity of each fragment decays because of air resistance or drag. Fragment velocity decreases more rapidly with distance as the fragment weight decreases. For an assumed initial fragment velocity of 1,825 meters per second, a five-grain (0.324 gram) fragment would lose half its initial velocity after traveling 11.25 meters, whereas a 500-grain (32.4 grams) fragment would travel 53.34 meters before losing half its velocity.

Fragments of a warhead travel outward in a nearly perpendicular direction to the surface of the casing; for a cylindrical warhead, there is a 7°–10° lead angle. Figure 18-13 shows a typical fragmentation pattern. The tail and nose spray are frequently referred to separately as the 45° cone, which is an area of less dense fragmentation. If this payload were detonated in flight, the dense side spray would have a slight forward thrust with an increased velocity equal to weapon flight velocity.

The angle of the side spray in figure 18-13 would be defined as "the beamwidth of this fragmenting payload." Fragment beamwidth is defined as "the angle covered by a useful density of fragments." Beamwidth is a function of warhead shape and the placement of the detonator(s) in the explosive charge.

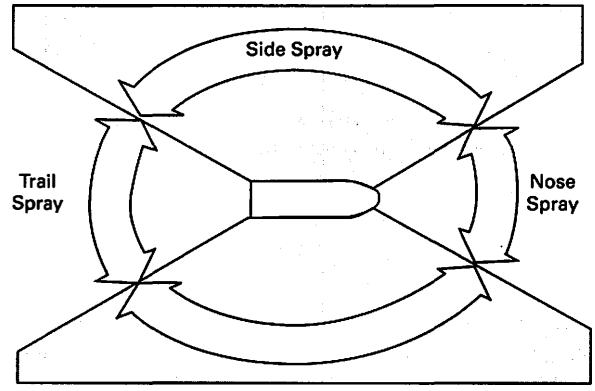


FIGURE 18-13 *Distribution of fragments*

The latest air target warheads are designed to emit a narrow spray of high-velocity fragments. This type of warhead, called an annular blast fragmentation (ABF) warhead, has a fragmentation pattern that propagates out in the form of a ring with tremendous destructive potential.

Another type of fragmentation warhead is the selectively aimable warhead (SAW). This "smart" warhead is designed to aim its fragment density at the target. This is accomplished by the fuzing system telling the warhead where the target is located and causing it to detonate so as to maximize the energy density on the target.

As a general rule, the vulnerability of some targets to damage caused by fragments from a warhead depends on the fragments' kinetic energy. The initial energy can be found from the Gurney analysis. Velocity as a function of range can be found from the drag equation. The kinetic energy equation relates mass of the fragment to its velocity at range *S*:

$$KE = \frac{1}{2} m V_s^2 \qquad (18\text{-}14)$$

where:

KE = kinetic energy (J)
m = mass of individual fragment (kg)
V_S = velocity at range S (m/s)

Rough approximations can be made to determine the amount of kinetic energy required to penetrate different targets. Based on typical ballistics numbers, 100 J is the minimum lethal kinetic energy against personnel. This is roughly equivalent to a .22-caliber long bullet (40 grains) from a rifle at 1,000 fps.

The next level of damage is about 1,000 J, which corresponds to a .357-caliber jacketed soft-point (158 grains) bullet at 1,400 fps. This is fairly lethal, depending on where it hits unprotected personnel.

Lastly, around 4,000 J is sufficient to penetrate body armor. This is equivalent to a 7.62 mm full metal jacket or a .30–06 armor-piercing bullet (166 grains) at 2,750 fps.

Aircraft are generally constructed of light metals. Giving a conservative estimate, aircraft skin can be treated as the equivalent of body armor. Thus it requires about 4,000 J to penetrate aircraft skin.

It is probably unwise to assume that an armored vehicle could be stopped by a fragmentation warhead. Any substantial amount of armor would require a specialized projectile. For example, light armor vehicles use saboted shells, which have a .50-caliber outside shoe (sabot) containing a .30-caliber shell (penetrator) that is hardened and shaped to pierce armor. At almost 4,000 fps, this can penetrate 3/4 inch of steel. Armor up to about 15 inches can be pierced by special rounds that weigh more than 3.5 kilograms and travel at 700 m/s, making their kinetic energy about 850 kJ.

TABLE 18-9 *Sample damage criteria for fragmentation effects*

	Fragment Energy (kJ)		
Target	Light Damage ($P_{d/hit}$ = 0.1)	Moderate Damage ($P_{d/hit}$ = 0.5)	Heavy Damage ($P_{d/hit}$ = 0.9)
Personnel	0.1	1	4
Aircraft	4	10	20
Armored vehicle	10	500	1,000

As a crude rule of thumb, an estimate can be made that it takes about 10 kJ of kinetic energy per centimeter of steel in order to penetrate the steel. Table 18-9 is a summary of the damage criteria for targets vulnerable to fragmentation warheads.

Fragment trajectories will follow paths predicted by principles of external ballistics. For determining the effectiveness of almost all fragmenting munitions, the subsonic trajectory of the fragments can be ignored. As a result, the density of fragments in a given direction varies inversely as the square of the distance from the weapon. The number of hits on some unshielded target is proportional to the exposed projected area and inversely proportional to the square of the distance from the weapon ($1/R^2$). For an isotropic warhead:

$$N_{hits} = \frac{A_t N_0}{4\pi r^2} \qquad (18\text{-}15)$$

where:

N_{hits} = expected number of fragments hitting the target
N_0 = initial number of fragments from the warhead
A_t = frontal area of the target presented to the warhead
r = range of the target to the warhead

Because of the possibility of multiple hits by a fragmentation warhead, an estimation for P_d is found from:

$$P_d = 1 - (1 - P_{d/hit})^{N^{hits}} \quad \text{if } N_{hits} > 1 \qquad (18\text{-}16)$$

or

$$P_d = N_{hits} \times P_{d/hit} \quad \text{if } N_{hits} \leq 1 \qquad (18\text{-}17)$$

Continuous-Rod Warhead

A continuous-rod warhead is a special type of fragmentation round developed for use against aircraft. Early warhead experiments with short, straight, unconnected rods had shown that such rods could chop off propeller blades,

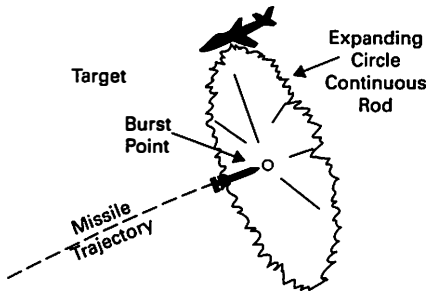


FIGURE 18-14 *Expansion of a continuous-rod warhead*

engine cylinders, and wings and, in general, inflict severe damage to a fighter aircraft.

However, rod warheads were ineffective against larger planes because the nature of most bomber aircraft structures permits a number of short cuts in their skin without lethal damage occurring. It was found, however, that long, continuous cuts would do considerable damage to a bomber; therefore, the continuous-rod warhead was developed.

Upon detonation, the continuous-rod payload expands rapidly into a ring pattern. The intent is to cause the connected rods during their expansion to strike the target and produce damage by a cutting action; see figure 18-14.

Each rod is connected end to end alternately and arranged in a bundle radially around the main charge. The burster is designed such that upon detonation, the explosive force will be distributed evenly along the length of the continuous-rod bundle. This is important in order to ensure that each rod will maintain its configuration and consequently result in uniform integrity of the expanding circle.

Figure 18-15 illustrates the arrangement of the bundle on a section of the main charge, and its accordionlike appearance as the section begins expanding.

The metal density of a normal fragmentation warhead attenuates inversely with the square of the distance ($1/R^2$). However, because it is nonisotropic, the metal density of a continuous-rod payload attenuates inversely as the distance from the point of detonation (1/R).

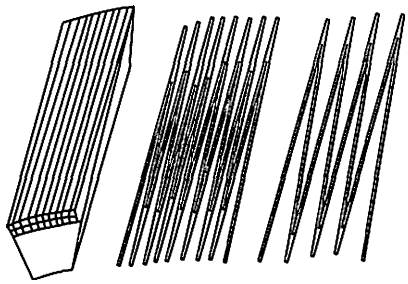


FIGURE 18-15 *Section of continuous-rod bundle*

To ensure that the rods stay connected at detonation, the maximum initial rod velocity is limited to the range of 1,050 to 1,150 meters per second. The initial fragment velocities of fragmentation warheads are in the range of 1,800 to 2,100 meters per second. Thus, in comparison, continuous-rod warheads cannot produce as much destructive energy potential as fragmentation warheads.

Shaped Charge Warheads

The first documented discovery in the United States of what is variously referred to as the shaped charge effect, the hollow charge effect, the cavity effect, or the Munroe effect dates back to the 1880s. Dr. Charles Munroe, while working at the Naval Torpedo Station at Newport, Rhode Island, in the 1880s, discovered that if a block of guncotton with letters countersunk into its surface was detonated with its lettered surface against a steel plate, the letters were indented into the surface of the steel.

The essential features of this effect were also observed in both Germany and Norway at approximately the same time, although no great use was made of it, and it was temporarily forgotten.

Shaped charge research became very important between 1935 and 1950 due to man-held antitank weapons. The major powers during World War II conducted extensive research and fielded shaped charge warheads capable of penetrating armor, such as the U.S. bazooka and Germany's panzerfaust.[iii]

A shaped charge warhead consists basically of an explosive-filled cylinder that is open in the front, with a hollow cavity lined with a thin layer of metal and glass or ceramic, usually conical hemispherical shaped, and backed on the convex side by explosive. A thin nose cone for aerodynamics covers the warhead, in addition to a fuze. The booster is located on the base of the warhead. See figure 18-16.

When a shaped charge warhead strikes a target, the fuze fires a length of detonation cord leading to the booster, which detonates the main charge from the rear. A detonation wave sweeps forward at detonation velocity and begins to collapse the metal cone liner at its apex. The collapse of the cone results in the formation and ejection of a continuous high-velocity molten jet of liner material.

The velocity of the tip of the jet is on the order of 8,500 meters per second, whereas the trailing end of the jet has a velocity on the order of 1,500 meters per second. This produces a velocity gradient that tends to stretch out or lengthen the jet until it separates into a column of fine particles.

The jet is then followed by a slug that consists of about 80% of the liner mass. The slug has a velocity on

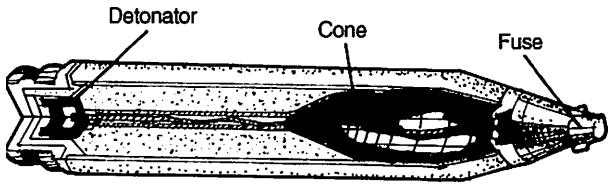


FIGURE 18-16 *Shaped charge warhead*

the order of 600 meters per second. This process is illustrated in figure 18-17.

When the energetic jet strikes a target of armor plate or mild steel, pressures in the range of 100–200 GPa[iv] are produced at the point of contact. This pressure produces stresses far above the yield strength of steel, and the target material flows like a fluid out of the path of the jet. It is important to note that the cavity produced is not due to a thermal effect but to the lateral displacement of the metal because of the immense pressures created.[v] This phenomenon is called hydrodynamic penetration.

There is so much radial momentum associated with the flow that the difference in diameter between the jet and the hole it produces depends on the characteristics of the target material. A larger diameter hole will be made in mild steel than in armor plate because the density and hardness of armor plate are greater. The depth of penetration into a very thick slab of mild steel will also be greater than that into homogeneous armor.

In general, the depth of penetration depends upon five factors:

1. Length of jet
2. Velocity of the jet tip
3. Density of the jet
4. Density of the target material
5. Jet precision (straight vs. divergent)

Longer jet length causes greater depth of penetration. Therefore, the greater the distance from the target to the base of the cone (known as the standoff distance), the better; see figure 18-18. This is true up to the point at which the jet particulates or breaks up, usually around six to eight cone diameters from the cone base. Particulation is a result of the velocity gradient in the jet, which stretches it out until it breaks up.

The depth of jet penetration can be approximated using the velocity of the jet tip and the density of the liner and target material. The velocity of the jet tip is given by[vi]:

$$V_0 = \frac{0.6002\, D\frac{C}{M}}{\cos\theta\left(2+\frac{C}{M}\right)} \qquad (18\text{-}18)$$

where:

V_0 = velocity of jet tip
D = velocity of detonation
C = mass of charge
M = mass of liner
θ = angle of shape charge

Using equation 18-18, the target jet density ratio, and the assumed minimum speed for penetration on the order

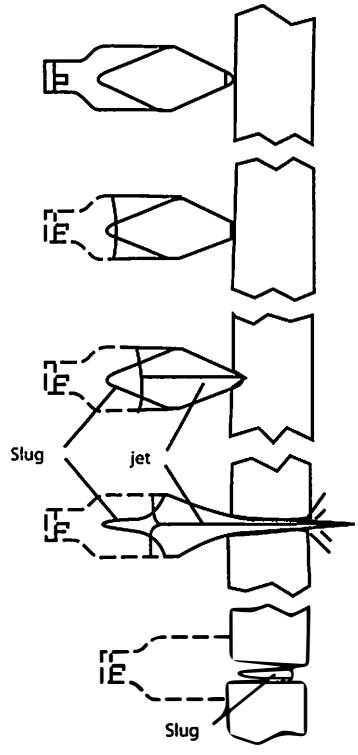


FIGURE 18-17 *Shaped charge sequence*

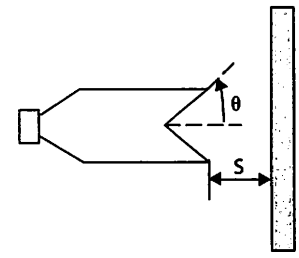


FIGURE 18-18 *Shaped charge definitions*

of 1,000 m/sec, the penetration of a solid (nonperforated) jet is given by[vii]:

$$P = S\left[\left(\frac{V_0}{\left(1+\sqrt{\frac{\rho_t}{\rho_j}}\right)(V_{\min})}\right)^{\sqrt{\frac{\rho_j}{\rho_t}}} - 1\right] \qquad (18\text{-}19)$$

where:

P = penetration distance
S = standoff distance ($0 \le S \le 3 \times$ diameter of charge)
ρ_t = density of target material
ρ_j = density of liner material
V_{min} = minimum speed to maintain jet cohesion ≈ 1,000 m/s

Jet precision refers to the straightness of the jet. If the jet is formed with some oscillation or wavy motion, then depth of penetration will be reduced. This is a function of the quality of the liner and the initial detonation location accuracy. In addition, the effectiveness of shaped charge warheads is reduced when they are caused to rotate. Degradation begins at ten revolutions per second.

This reduction in effectiveness occurs when angular velocity of the jet causes the centrifugal forces to exceed the yield strength of the material and tear the jet apart. Thus, spin-stabilized projectiles generally cannot use shaped charge warheads unless special spin compensation is applied to the jet through the use of fluted liners or special metallurgical spin compensation techniques used during warhead fabrication.[viii]

The effectiveness of a shaped charge payload is independent of the striking velocity of the warhead. In fact, the velocity of the warhead must be taken into consideration to ensure that detonation of the payload occurs at the instant of optimum standoff distance. The jet can then effectively penetrate the target.

Various theories have been advanced concerning the action that takes place within a tank when it is perforated by the jet of a shaped charge. Tests show that for shaped charge warheads weighing 15 pounds or less, there is no appreciable blast effect, pressure rise, or temperature rise inside the tank. The jet does not spread out in a cone-shaped spray as it exits from the armor plate. Damage incurred is mostly a function of the jet and material from the target armor detached off the rear face.

If the jet perforates the armor plate with enough energy, it continues along its original path, causing ammunition, fuel, and so on, to be set afire and crew members killed. Damage is actually produced by the tiny, hot, high-velocity fragments of the liner metal making up the jet. The action of the target armor detaching and the material joining with the shaped charge jet is known as spalling. The extent of spalling is a function of the amount of explosive in the payload and the quality of the target armor.

Special-Purpose Warheads

There are means of attacking targets other than with blast, fragmentation, shaped charge, or continuous-rod payloads. Several types of payloads are more specialized in nature, designed to perform a specific function. A few of these will be described.

Armor Piercing

Because of extensive innovative advances in tank armor, shaped charge warheads have grown in diameter, and other types of warheads have been developed. Two of the most common are the kinetic energy penetrator and the plastic armor defeat charge.

The kinetic energy warhead employs a very heavy, hard, metal-core penetrator that travels at extremely high velocities. The penetrator is fin stabilized and uses a discarding sabot to increase its size to fit the gun barrel diameter when fired. The armor plate is thus defeated by either ductile or plastic flow failure, or by shearing or plugging such as a cookie cutter would do. The shape of the penetrator tip on this weapon or any other weapon is the determining factor. These are also called armor-piercing, fin-stabilized, discarding-sabot (APFSDS) rounds, as shown in figure 18-19.

The high-explosive, plastic defeat mechanism uses a high-explosive plastic filler in a shell that deforms on impact to effectively put a large glob or cone of plastic explosive against the side of the armor. The timing of the base detonator is critical for maximum effect.

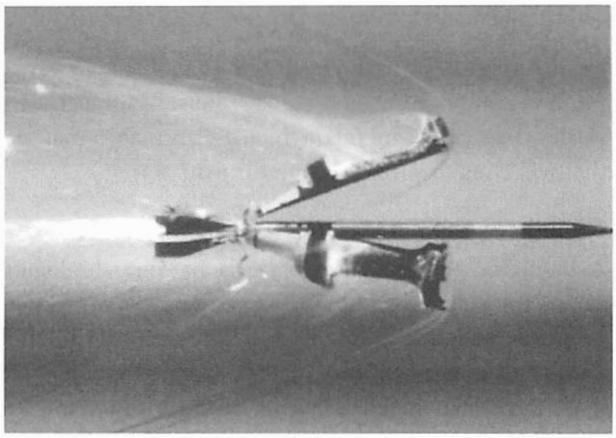


FIGURE 18-19 *120 millimeter kinetic energy antitank projectile*[ix]

The armor is not actually penetrated, but extensive spalling is caused on the opposite side of the armor. This warhead is limited to lighter armor than are the shaped-charge or armor-piercing kinetic energy warheads.

Fuel-Air-Explosive [FAE]

Fuel-air-explosives function by dispersing an aerosol cloud of fuel that is then ignited to produce a violent explosion. The rapidly expanding overpressure wave flattens all objects within close proximity of the epicenter. There are dramatic differences between FAE and chemical high explosives.

Chemical explosives exhibit a high explosive blast overpressure over a relatively short duration, whereas an FAE blast overpressure is lower, and its duration is much longer, thereby producing a more enhanced blast impulse. FAE represents the military application of vapor cloud and dust explosion accidents that have long bedeviled a variety of industries and is very effective against blast-vulnerable targets.

Thermal Warheads

The purpose of thermal warheads is to start fires. These payloads may employ chemical energy to kindle fires with subsequent uncontrollable conflagrations, or nuclear energy to produce direct thermal destruction as well as subsequent fires. Thermal payloads of the chemical type may be referred to as incendiary or firebombs. Many area targets are more effectively attacked by fire than by blast or fragmentation.

Incendiary material such as napalm is made simply by combining a conventional fuel (gasoline or JP-5 jet fuel) with an thickener to create an extremely flammable gel.

Another type of thermal weapon is made by combining metals. One example is thermite, which is a mixture of aluminum and iron oxide that when combined and starts to burn, produces molten iron. A thermite hand grenade spreads pieces of molten iron that in turn create secondary fires.

These warheads are used to burn the area surrounding the point of impact, which can be an effective way to remove ground cover as well as troops. The incendiary warhead usually contains a highly flammable material that will cling to surfaces when thrown.

Biological and Chemical Warheads

A biological warhead uses bacteria or other biological agents for accomplishing its purpose of causing sickness or death and is of extreme strategic importance since it is capable of destroying life without damaging buildings or materials. The poisoning of water supplies is probably the single most efficient way of destroying enemy personnel. The war potential of the enemy, such as guns, missile launching sites, and so on, are thus left intact and at the disposal of the attacker.

The biological agent may be chosen so that it causes only temporary disability rather than death to enemy personnel, thereby making it relatively simple to capture an enemy installation. A small explosive charge placed in a biological payload is useful in the dispersion of biological agents.

A chemical warhead payload is designed to expel poisonous substances and thus produce personnel casualties. Binary warheads are stored with two inert subsections. When properly fuzed, they combine to form a lethal payload. Chapter 19 will address this topic in detail.

Radiation Warheads

All nuclear weapons emit radiation. However, an enhanced radiation weapon can be designed to maximize this effect. Chapter 19 will address this topic in greater detail.

Pyrotechnic Warheads

Pyrotechnics are typically employed for signaling, illuminating, or marking targets. In their simplest form they are handheld devices. Some examples of more elaborate warhead payloads are as follows:

1. *Illumination.* These warheads usually contain a flare or a magnesium candle as the payload, which is expelled by a small charge and is parachuted to the ground. During its descent, the flare is kindled.

   The illuminating warhead is thus of great usefulness during night attacks in pointing out enemy fortifications. Illumination projectiles are used with great effectiveness in shore bombardment. Illuminating warheads are also used as aircraft flares and flare rockets to assist in the attack of ground targets and submarines.

   Because these flares are difficult to extinguish if accidentally ignited, extreme caution in their handling is required. An illuminating projectile is shown in figure 18-20.
2. *Smoke.* These warheads are used primarily to screen troop movements and play a vital role in

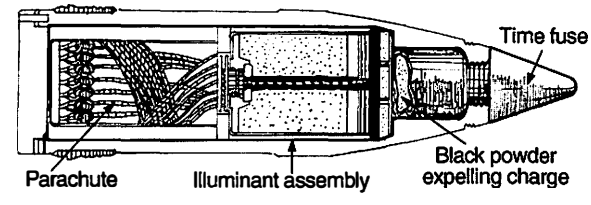


FIGURE 18-20 *Typical illuminating projectile*

battlefield tactics. A black powder charge ignites and expels canisters that may be designed to emit white, yellow, red, green, or violet smoke.

3. *Markers.* White phosphorus is commonly employed as a payload to mark the position of the enemy. It can be very dangerous, especially in heavy concentrations. The material can self-ignite in air, cannot be extinguished by water, and will rekindle upon subsequent exposure to air. Body contact can produce serious burns. Copper sulphate prevents its reignition.

Hard Target Penetrator

For targets buried deep underground or protected by massive layers of concrete, a warhead called a hard target penetrator is used. This weapon uses casings made of a dense metal or contains dense metal ballast to gain momentum and penetrate deeply into the target.

The fuzing is a special accelerometer-based electronic fuze that allows control of the detonation point by layer counting of distance or time. The accelerometer senses g-loads on the bomb due to deceleration as it penetrates through to the target. The fuze can distinguish between earth, concrete, rock, and air. These weapons have penetrated more than 20 feet of reinforced concrete, and are capable of penetrating more than 100 feet of earth.[x]

Antipersonnel Warheads

Such warheads are designed to kill or maim personnel or to damage material enough to render it inoperable. In the area of field artillery, the flechette or beehive round is an example of an antipersonnel warhead. The payload in this projectile consists of 8,000 steel-wire, fin-stabilized darts.

Upon detonation, the darts, or flechettes, are sprayed radially from the point of detonation, normally within 60 feet of the ground. These warheads are extremely effective against personnel in the open or in dense foliage.

Chaff Warheads

Chaff may be employed to decoy enemy weapons or blind enemy radar. The payload typically consists of metal-coated fiberglass strands cut in lengths determined by the wavelength of the RF energy to be countered. Chaff may be dispensed in a variety of warheads including projectiles and rockets.

Cluster Bomb Units (CBU)

CBUs are air-delivered canisters containing hundreds of small bomblets for use against a variety of targets such as personnel, armored vehicles, and ships; see figure 18-21. Once in the air, the canisters open, spreading the bomblets out in a wide pattern. The advantage of this type of warhead is that it gives a wide area of coverage, which allows for a greater margin of error in delivery.

There are many types of specialized bomblets; some contain antipersonnel darts; some contain land mines; some have shaped charge warheads for wide dispersion antitank roles; and still others have carbon graphite filaments to disrupt power generation facilities.[xi]

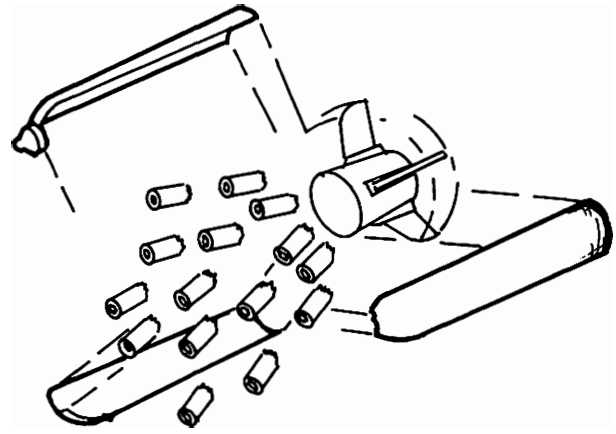


FIGURE 18-21 *Cluster bomb separation*

Psychological

This type of warhead does not carry a lethal or destructive agent, but instead carries material designed to have a psychological effect on the enemy rather than cause actual damage. Payloads such as propaganda leaflets, mysterious objects that appear to be dangerous, and inert or dummy warheads are examples. A leaflet bomb is shown in figure 18-22.

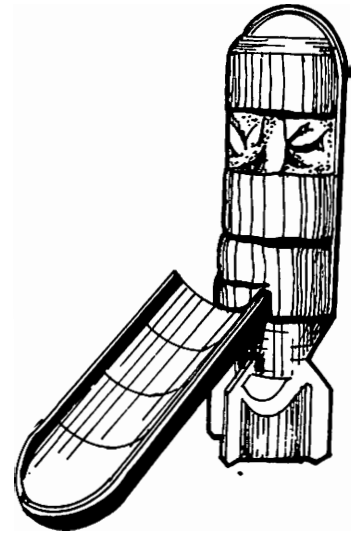


FIGURE 18-22 *Leaflet bomb*

Underwater Warheads

The mechanism of an underwater blast presents some interesting phenomena associated with a denser medium than air. An underwater explosion creates a cavity filled with high-pressure gas, which pushes the water out radially against the opposing external hydrostatic pressure.

At the moment of explosion, a certain amount of gas is instantaneously generated at high pressure and temperature, creating a bubble. In addition, heat causes a certain amount of water to vaporize, adding to the volume of the bubble. This action immediately begins to force the water in contact with the blast front in an outward direction. Potential energy initially possessed by the gas bubble by virtue of its pressure is thus gradually transferred to the water as kinetic energy.

The inertia of the water causes the bubble to overshoot the point at which its internal pressure is equal to the external pressure of the water. The bubble then becomes rarefied, and its radial motion is brought to rest. External pressure now compresses the rarefied bubble. Again, the equilibrium configuration is overshot, and since by hypothesis there has been no loss of energy, the bubble comes to rest at the same pressure and volume as at the moment of explosion; in practice, energy is lost by acoustic and heat radiation.

The bubble of compressed gas then expands again, and the cycle is repeated, resulting in a pulsating bubble of gas slowly rising to the surface, with each expansion of the bubble creating a shock wave. Approximately 90% of the bubble's energy is dissipated after the first expansion and contraction. This phenomenon explains how an underwater explosion appears to be followed by other explosions. The time interval of the energy being returned to the bubble (the period of pulsations) varies with the intensity of the initial explosion.

The rapid expansion of the gas bubble formed by an underwater explosion results in a shock wave that is sent out through the water in all directions. The shock wave is similar in general form to a shock wave in air, although it differs in detail.

Just as in air, there is a sharp rise in overpressure at the shock front. However, in water, the peak overpressure does not fall off as rapidly with distance as it does in air. Hence, peak values in water are much higher than those at the same distance from an equal explosion in air.

The velocity of sound in water is nearly 1 mile per second, almost five times as great as in air. Consequently, the duration of the shock wave developed is shorter than in air.

The close proximity of the upper and lower boundaries (the water surface and the ocean floor) between which the shock wave is forced to travel causes complex shock wave patterns as a result of reflection and rarefaction. Also, in addition to the initial shock wave that results from the initial gas bubble expansion, subsequent shock waves are produced by bubble pulsation. The pulsating shock waves are of lower magnitude and of longer duration than the initial shock wave.

The peak pressure expands outward from the center of the explosion at a velocity of 1,550 m/sec, slowing to approximately 1,450 m/sec. Pressure as a function of range and warhead size can be found using the following thumb rule[xii]:

$$P_{max} = 53.1\left(\frac{W^{1/3}}{r}\right)^{1.13} \qquad (18\text{-}20)$$

where:

P_{max} = peak pressure in mega Pascals
W = weight of the warhead in kilograms of TNT
r = range from blast in meters

Another interesting phenomenon of an underwater blast is surface cutoff. At the surface, the shock wave moving through the water meets a much less dense medium. As a result, a reflected wave is sent back into the water, but this is a rarefaction or suction wave.

At a point below the surface, the combination of the reflected suction wave and the direct incident wave produces a sharp decrease in the water shock pressure. This is surface cutoff. Variation of the shock overpressure with time after the explosion at a point underwater not too far from the surface is illustrated in figure 18-23.

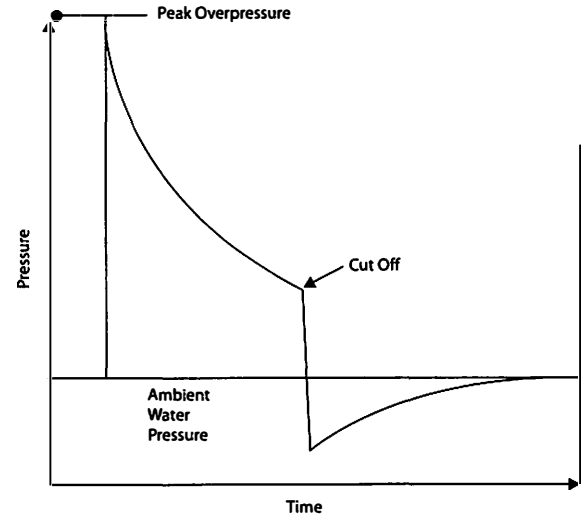


FIGURE 18-23 *Pressure versus time for an underwater explosion (near the surface)*

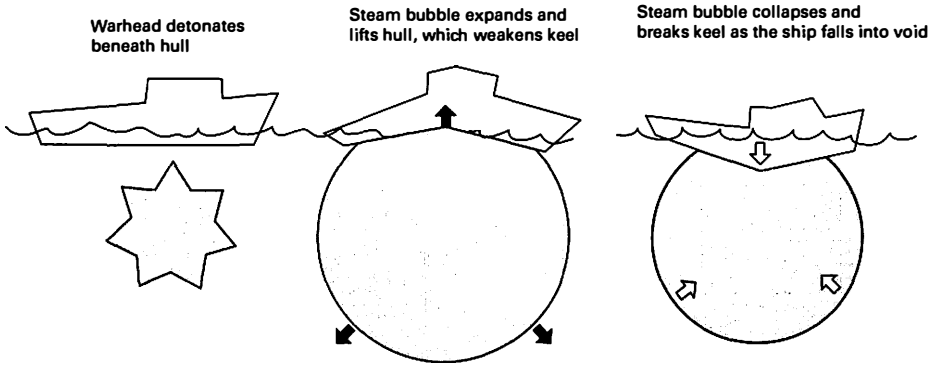


FIGURE 18-24 *Beneath keel explosion.*

After the lapse of a short interval, which is the amount of time required for the shock wave to travel from the explosion to the given location, the overpressure rises suddenly due to the arrival of the shock front. Then, for a period of time, the pressure decreases steadily, as in air.

Soon thereafter, the arrival of the reflected suction wave from the surface causes the pressure to drop sharply, even below the normal hydrostatic pressure of the water. This negative pressure phase is of short duration and can result in a decrease in the extent of damage sustained by the target.

The time interval between the arrival of the direct shock wave at a particular location in the water and that of the cutoff, signaling the arrival of the reflected wave, depends upon the depth of burst, the depth of the target, and the distance from the burst point to the target.

It can generally be said that a depth bomb should be detonated at or below the target and that a target is less vulnerable near the surface.

Targets exposed to underwater shock waves will be subject to diffraction and drag loading just as they are in the air. There will be very little translation of targets because the shock wave is of short duration, and objects in motion will quickly be brought to rest by the drag force of the surrounding water.

Diffraction loading applies, but targets that are in the water tend to be well designed to withstand this type of loading. Therefore, very little damage is done directly by these mechanisms.

Considerable damage can still occur due to internal equipment failure and missile hazards—for instance, things thrown through the air inside the ship. In order to be the most effective, underwater warheads utilize the interaction of the steam void created by the explosion with the hull of the target. There are two distinct cases, surface and submerged bodies.[xiii]

When a warhead is detonated at close range beneath a ship, the steam bubble initially lifts the ship upward from the middle; see figure 18-24. This tends to weaken the ship's keel. After the steam bubble has reached its maximum volume, the surrounding water pressure will collapse it.

The ship then falls into the void, still supported on its ends. The keel will then break under the ship's own weight. The compression of the steam void will raise the temperature and the bubble will oscillate a few times. The ship may be destroyed during the subsequent oscillations even if it manages to survive the first.[xiv]

When used against a submerged target, the greater ambient pressure and buoyancy of the surrounding water preclude effective uses of this technique. However, the interaction of the steam bubble and the submerged hull can be exploited to severely damage or potentially breach the submarine's pressure hull.

The process begins with a close-range underwater explosion. If the steam bubble contacts the hull, it will attach itself. In this location, the oscillation of the steam bubble will cause a cyclic stress to the submerged hull and weaken or breach it. If a submarine's pressure hull is compromised, it is unlikely to survive. This is illustrated in figure 18-25.[xv]

Mine Warheads

Mine warheads use the underwater blast principles described earlier to inflict damage to the target ship or submarine. The damage energy transmitted is approximately equally divided between the initial shock wave and the expanding gas bubble. If the target is straddling the gas bubble, then it will have unequal support and may be broken in two.

As the detonation depth increases, particularly in excess of 180 feet, the effect of the gas bubble causing

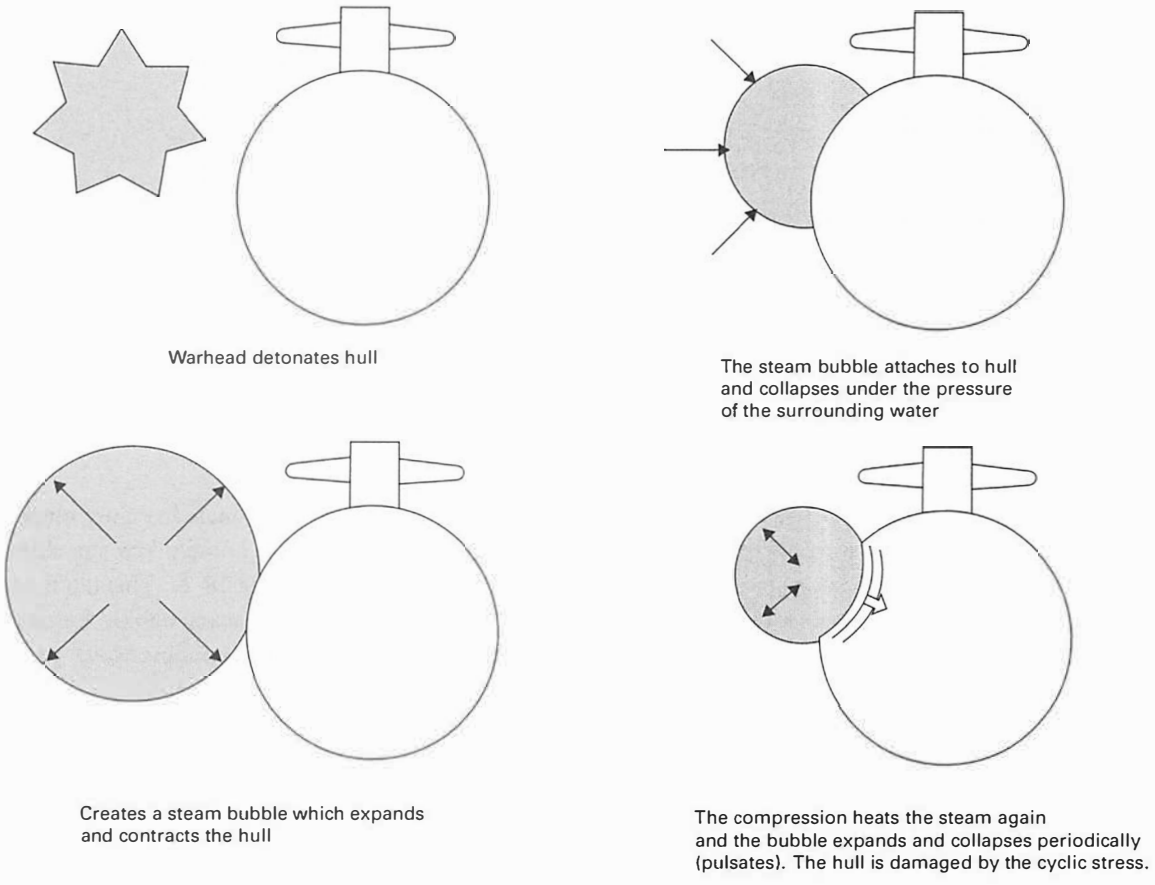


FIGURE 18-25 *Explosion near a submarine hull*

damage is greatly diminished; therefore, bottom mines are rarely used in waters exceeding 180–200 feet.

Torpedo Warheads

Torpedo warheads must be capable of damaging both ships and submarines. Homing in on the screws can achieve a mobility kill. Detonation under the keel at amidships can cause the severe gas bubble damage mentioned above, and if the depth is less than 300 feet, the reflected shock wave can substantially increase the damage effects.

Torpedoes that actually impact the hull of a ship or submarine must overcome the double hull/void structure of the vessel. As such, deep-diving submarines with especially thick hulls require highly specialized warheads.

---

[i] *FAS* (http://www.fas.org/man/dod-101/sys/land/ docs/warheads.pdf)

[ii] Hall, 290

[iii] Walters, 14

[iv] Walters, 14

[v] Walters, 15

[vi] Walters, 61 and 64

[vii] Walters, 141

[viii] Walters, 35

[ix] Courtesy General Dynamics Ordnance

[x] *FAS* (http://fas.org/man/dod-101/sys/smart/index. html)

[xi] *FAS* (http://fas.org/man/dod-101/sys/dumb/index. html)

[xii] Keevin

[xiii] Hall, 300

[xiv] Hall, 300

[xv] Hall, 300

19 Weapons of Mass Destruction

Introduction

Chemical, biological, and radiological (CBR) weapons (often referred to in older literature as nuclear, biological, and chemical (NBC) weapons) collectively are referred to as weapons of mass destruction (WMD) since they are capable of large amounts of destruction or can be used against large numbers of personnel. In this uncertain world, these weapons pose a significant threat to people of all nations. This threat exists from both nations and terrorist groups who see such weapons as an attractive means to advance their agendas.

This chapter is not designed to be a single source reference for WMD; rather, it is devoted to providing the student with a fundamental understanding behind these weapons and their catastrophic, devastating potential.

Nuclear Weapons

Although the threat of a nuclear exchange between superpowers has diminished, the likelihood of a localized exchange of nuclear weapons has increased with the continued proliferation of nuclear technology and the increasing chance that such weapons might be acquired by either a rogue state or a terrorist organization. Although total nuclear war with the utilization of all available nuclear weapons could result in the complete devastation of the nations involved, situations smaller than total nuclear war are also possible in which nuclear weapons are employed in limited numbers or for a limited time, along with conventional weapons.

Under such circumstances, effective military operations would be expected to continue, and the military officer should have a basic understanding of the effects of such weapons.

Principles of Operation

Nuclear warheads may be classified as either fission or fusion, depending on the nature of the reaction's initiation. Both apply the same general principle of changing the configuration of the atomic nucleus, but how the nucleus's configuration is changed is the fundamental difference between fission and fusion.

Fusion combines two or more light nuclei into a nucleus of a heavier atom, whereas fission splits a heavy nucleus into two or more lighter nuclei. In either case, significant energy is released that can be harnessed for either peaceful—generation of nuclear power—or destructive purposes.

The nucleus is held together by the strong nuclear force, which is an attractive force between nucleons (protons and neutrons). Changing the configuration of the nucleus usually changes the binding energy of the nucleus.

Binding energy is a measure of the energy needed to separate the nucleus into its constituent protons and neutrons or, conversely, the energy liberated when protons and neutrons combine to form an atom. When binding energy increases, the energy stored in the nucleus is reduced. The excess energy is then liberated from the nucleus.

These forces have orders of magnitude that are stronger than the normal electromagnetic force in the outer shells of the atoms that holds molecules together. Therefore, the magnitude of the nuclear binding energy is much greater than the magnitude of the molecular binding energy.

For example, about 24 eV (electron volts) are required to remove one of the two electrons in a helium atom, and about 54 eV are needed to remove the second electron; therefore, the total energy needed to remove both of the electrons would be 78 eV. By contrast, about 28 million electron volts (MeV) are required to separate the two protons and two neutrons in the helium nucleus from each other.

The nuclear binding energy of helium is larger than its molecular binding energy by a factor of 3.6×10^5. Since the magnitude of the nuclear binding energy is so much greater, the amount of energy released in nuclear reactions is also much greater; consequently, nuclear explosions are much more powerful than conventional explosives when the equivalent quantities of material are used.

Fission vs. Fusion

Although binding energy increases with the addition of more and more nucleons, a point is reached where the nuclear attractive forces among nucleons begin to decrease due to the increasing size of the nucleus. If binding energy per nucleon is plotted for increasing nuclei size, figure 19-1 is produced.

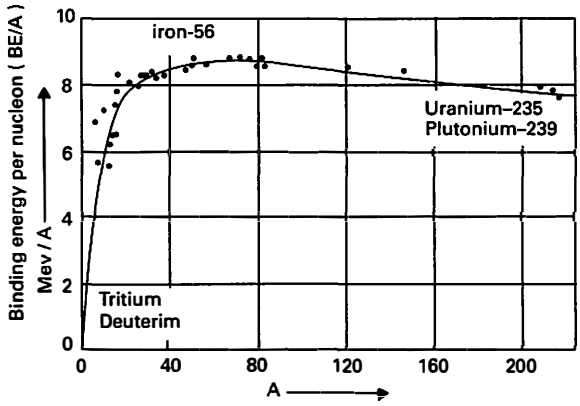


FIGURE 19-1 *Binding energy (BE) per nucleon curve*

Binding energy per nucleon reaches its peak at about 56 nucleons, corresponding to the element iron (Fe). For elements with atomic numbers less than 56, it is theoretically possible for them to undergo fusion reactions to release energy; however, from a practical standpoint, the fusion of deuterium (D or ${}_1^2\mathrm{H}$) with tritium (T or ${}_1^3\mathrm{H}$) is the only militarily useful fusion reaction.

For fission, the energetically best situation occurs when the atomic number is very high. For very large nuclei, the process occurs spontaneously with regularity. Again, the process has diminishing returns the closer the parent nucleus is to the peak of the binding energy per nucleon curve.

The most commonly used fuels for fission warheads include ^{235}U and ^{239}Pu. When ^{235}U undergoes fission, it typically releases about 210 MeV of energy. We can estimate this by assuming that the nucleus splits into two equal parts, each with about 118 nucleons.

Before fission, the ^{235}U nucleus is held together with about 7.5 MeV per nucleon or 1,763 MeV total. After the split, the two parts have about 8.5 MeV per nucleon, or 2,000 MeV. The difference of 237 MeV is released in the reaction.

A single fusion reaction, however, actually produces less energy than its fission counterpart. The fusion of a deuterium (D or ${}_1^2\mathrm{H}$) nucleus (one proton and one neutron) with a tritium (T or ${}_1^3\mathrm{H}$) nucleus (one proton and two neutrons) would yield a single helium (${}_2^4\mathrm{He}$) nucleus, an extra neutron, ${}_0^1n$, and energy, as in the following formula:

$$ {}_1^2\mathrm{H} + {}_1^3\mathrm{H} \rightarrow {}_2^4\mathrm{He} + {}_0^1n + 18\ \mathrm{MeV} \qquad (9\text{-}1) $$

The released energy, in this case 18 MeV, equals the difference in binding energy before and after the reaction:

$$ \begin{aligned} \text{BE (before)} &= -2\ \text{MeV (D)} + -8\ \text{MeV (T)} \\ \text{BE (after)} &= -28\ \text{MeV (helium)} \end{aligned} $$

Therefore: (19-2)

$$ \begin{aligned} \text{energy released} &= -10\ \text{MeV} - (-28\ \text{MeV}) \\ &= +18\ \text{MeV} \end{aligned} $$

If every nucleus in 1 kilogram of a deuterium-tritium mixture underwent this fusion reaction, about 3.5×10^{14} Joules of energy would be released:

$$ (1{,}000\,\mathrm{g\ D\text{-}T}) \times \left(\frac{1\,\mathrm{mol}}{5\,\mathrm{g}}\right) \times \left(\frac{6.02 \times 10^{23}\ \mathrm{nuclei}}{1\,\mathrm{mol}}\right) \times \left(\frac{1.8 \times 10^{7}\,\mathrm{eV}}{\mathrm{nuclei}}\right) \times \left(\frac{1.6 \times 10^{-19}\,\mathrm{J}}{1\,\mathrm{eV}}\right) = 3.5 \times 10^{14}\,\mathrm{J} \qquad (19\text{-}3) $$

By comparison, the same mass of TNT would release only 4.680×10^6 J. The power index of 1 kilogram of D-T is approximately 7.5 billion times the strength of TNT:

$$ \mathrm{PI} = \left(\frac{3.5 \times 10^{14}\,\mathrm{J}}{4.680 \times 10^{6}\,\mathrm{J}}\right) \times 100 = 7.5 \times 10^{9}\% \qquad (19\text{-}4) $$

Although fusion actually liberates less energy per atomic reaction than fission, the larger number of atoms present in equivalent weights of fusionable versus fissionable material results in a significantly larger explosion.

Fission Warheads

During the initial design and research leading to the development of the atomic bomb, conditions had to be

created that would make a fission reaction occur with extreme reliability and rapidity—that is, a complete reaction occurring in 1 microsecond. Certain naturally occurring isotopes, such as ^{238}U and ^{239}Pu, were known to undergo spontaneous fission, a process of decomposition into smaller nuclei under natural conditions.

The slow reaction rates of the natural decay process must be increased by introducing additional neutrons into the material while also enriching naturally occurring uranium with the ^{235}U isotope. Natural uranium deposits are only approximately 0.7% ^{235}U; the remainder is a highly stable isotope, ^{238}U.

On average, a fission reaction of ^{235}U produces 2.54 neutrons, which may induce additional fissions within the fissile material, as shown in figure 19-2. Neutron generation across subsequent generations of neutron production is referred to as the neutron life cycle, and the fissions that occur are referred to as a chain reaction.

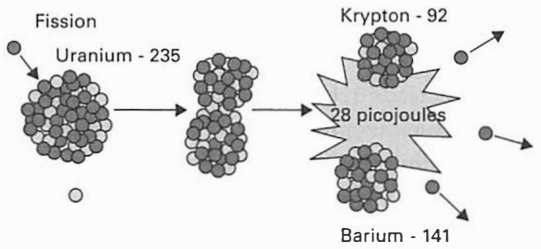


FIGURE 19-2 *Typical fission process of ^{235}U*

Supercritical mass If the fissile material is sufficiently large or densely packed such that neutrons produced in a subsequent generation are greater than the neutrons produced in the previous generation, a supercritical chain reaction is produced, and the neutrons and subsequent fissions increase at an exponential rate. This is the principle behind a fission weapon.

Critical mass A chain reaction that produces a neutron population that is essentially constant from one generation to the next is referred to as critical, and it is the basis for electrical generation from nuclear power plants.

Sub-critical mass A neutron population that decreases over subsequent generations and produces less than one neutron to carry on the fission process is referred to as a subcritical mass.

The detonation of a fission warhead begins with the creation of the conditions necessary to produce a supercritical chain reaction within the fissile material and an ever-increasing number of fissions until either the entire mass is consumed or the conditions within the fuel change, stopping the fission process. This depends on the specific fissile material selected as the fuel, the physical arrangement of the weapon components (such as the shape and density of the fissile material), and the temperature and pressure of the material.

In order to "kick-start" the reaction, a neutron source is typically implanted within the weapon to provide a high number of neutrons to begin the supercritical chain reaction when desired. The reaction proceeds in an uncontrolled, rapid manner, in about 1×10^{-8} second, and the result is a tremendous energy release that generates incredible heat, radiation, and a pressure shock wave of immense destructive power. In order to properly safeguard the weapon from premature detonation, the fissile material is kept in a subcritical state until detonation is initiated.

Gun Design Warhead

The uncomplicated gun-type warhead was chosen as the design for the first atomic weapon. Its features are shown in figure 19-3. Beginning with a subcritical mass of ^{235}U separated into two halves to prevent inadvertent detonation from spontaneous fission, upon initiation the weapon quickly drives the two subcritical masses together into a dense, compact mass while simultaneously injecting neutrons into the newly formed supercritical mass. The yield of the "Little Boy" weapon that was used at Hiroshima was approximately 15 kilotons (kT) of TNT.

The efficiency of this weapon was very low. Although putting the two fissile hemispheres together created the supercritical mass, the device expanded as soon as detonation began. As the fissile material came apart, density became subcritical and the chain reaction stopped, resulting in low efficiency and the wasting of fissile material. Additionally, the remaining fissile material itself became an additional hazard to personnel as radioactive fallout.

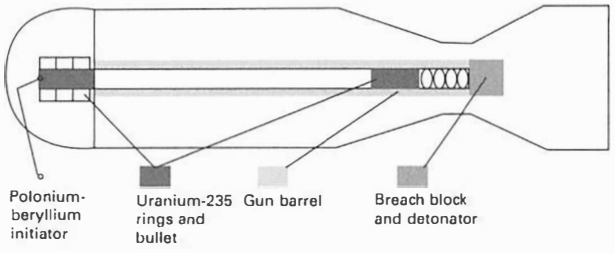


FIGURE 19-3 *Gun-type fission warhead "Little Boy"*

Implosive Warhead

Two technical problems existed with the gun-type weapon that needed to be addressed: excessive neutron leakage and premature ending of the chain reaction. Neutron leakage refers to neutrons that leave the body of the fissile material without reacting with the fuel. Excessive leakage decreases the overall neutron population and slows the rate of growth in the chain reaction. Surrounding the warhead with a neutron reflector, or tamper, reduces neutron leakage.

The terms *tamper* and *reflector* are generally used interchangeably, but have subtle differences. Whereas a reflecting material helps to scatter leaking neutrons back toward the fissile material, thereby improving efficiency by giving those neutrons that otherwise would have escaped another chance to produce fissions, the tamper actually serves a second purpose.

A tamper, usually beryllium-9 or uranium-238, additionally helps to confine the expanding fission material upon detonation, giving it a slightly longer time frame to fission before expansion stops the fission process altogether. By reducing the expansion rate of the fissile material, the tamper allows additional fissions to occur, which increases the efficiency and yield of the warhead while also substantially reducing the mass of fissile material needed to achieve a given yield.

The second problem to be solved was the possibility of the reaction initiating while the "bullet" was still some distance from the other half of the fissile material. If the bullet was not accelerated fast enough toward the other half of the material, a premature explosion could result in a very low-yield explosion called a "fizzle."

To overcome this problem, engineers designed an implosion weapon that allows the chain reaction to be sustained for a longer period of time. Fissionable material at subcritical density is encased inside a sphere of high-explosive lenses. Upon detonation of the high-explosive shell, the fissile material is then compressed to supercritical density, thereby initiating the nuclear explosion.

The fissile material continues to compress for a short time, and the additional compression increases the efficiency of the weapon tremendously. Eventually, the material is forced outward due to the pressure expansion from the nuclear explosion and reaches a point at which it is no longer a supercritical mass, thus ending the chain reaction.

The improved implosion-type warhead shown in figure 19-4 was developed in tandem with the gun-type warhead and implemented in the "Fat Man" bomb used on Nagasaki. This weapon had a yield estimated of 21 kT.

The use of thick tampers and high-energy explosives in modern warhead has reduced the required amount of fissile material to a small fraction of the amount needed in older designs.

Fusion Warheads

Fusion warheads require conditions that make fusion reactions overwhelmingly probable with extreme reliability and rapidity. Combining nuclei requires conditions of extremely high temperatures and/or pressure, such as in the sun. To create these conditions in a weapon, fission detonations must be used to trigger the fusion reaction. Hence, in every fusion bomb there is a fission initiator. Gamma radiation, produced from the fission reaction and discussed later in this chapter, transfers the necessary energy to the fusion material.

The fission detonation produces significant pressure and temperature, initiating the fusion reaction in the

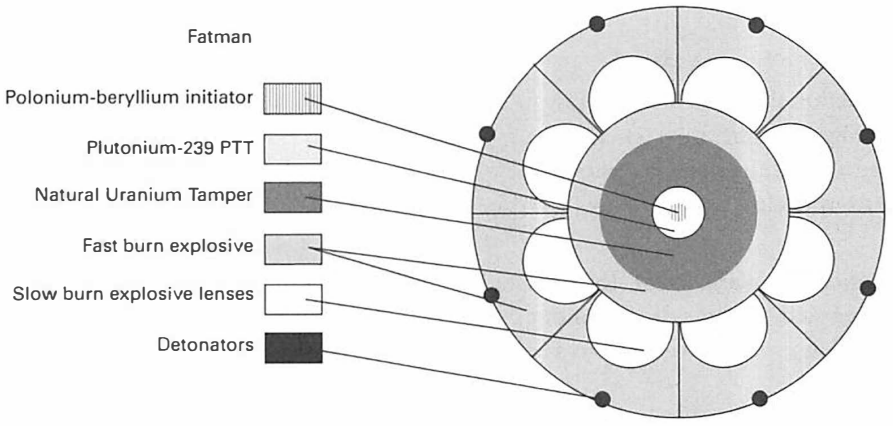


FIGURE 19-4 *Implosion warhead design*

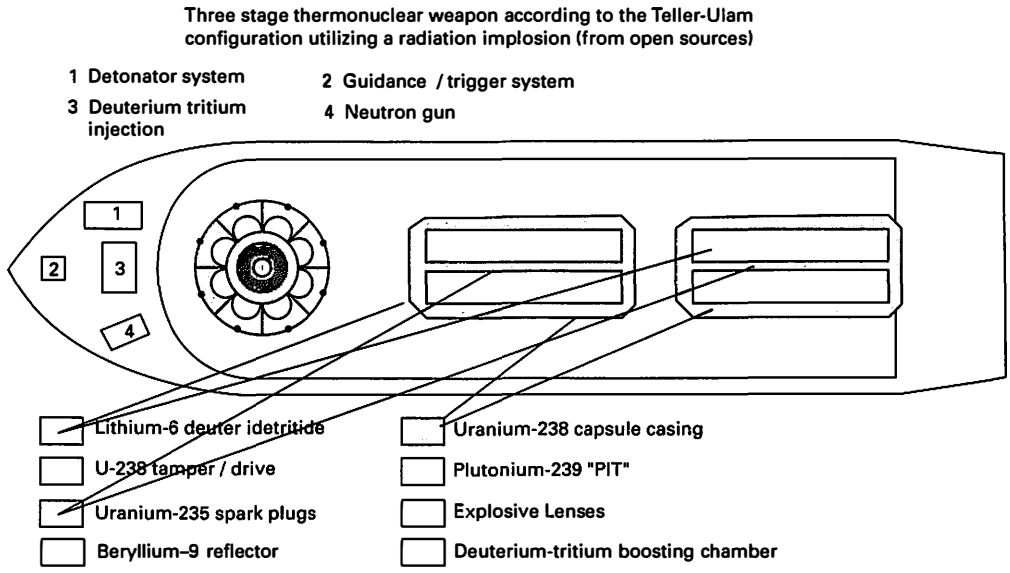


FIGURE 19-5 *Fusion warhead design*

fusion material. The pressure increase from a fission warhead can be more than 5×10^{12} N/m^2 (equivalent to about 7×10^8 psi), or 50 million times ambient pressure. This pressure compresses and subsequently heats the fusion material adiabatically.

The temperature of the fusion material increases by about the same ratio as the pressure increase, reaching more than 10^{10} K. At these high temperatures, the thermal motion of the atoms in the fusion material is sufficiently energetic that fusion becomes overwhelmingly probable; hence the name of this process is *thermonuclear fusion.*

All thermonuclear warheads are triggered by a fission warhead, called the primary or trigger. In order to focus the gamma rays and allow them to interact with the secondary fusion fuel, a good gamma-ray reflector surrounds the entire warhead. Very heavy atoms are best suited for this, and again ^{238}U makes an excellent gamma-ray reflector as well as provides the tamper for the primary fuel. Figure 19-5 shows a typical fusion warhead design.

Safety

Prevention of inadvertent detonation or intentional detonation by unauthorized individuals is paramount to nuclear weapon safety.

All warheads, including nuclear ones, have features to protect them from unintentional detonation. Since nuclear warheads are initiated by conventional high explosives, many safeguards are the same as those employed in safeguarding conventional warheads—namely, fuze installation just prior to deployment and physical and/or electronic interlocks that prevent arming until all safeguard conditions are met.

Furthermore, implosion fission warheads require exacting high-explosive configurations in order to symmetrically compress the fissile material into supercritical mass density. This requirement may also be engineered into the weapon design as a safety feature. If the high explosive is set off by some means other than the installed fuze, the precise geometry will not be obtained and no nuclear yield will be generated.

The prevention of unauthorized detonation is twofold: providing security for existing weapons and preventing the proliferation of nuclear weapons. Security measures taken to safeguard nuclear weapons are not generally different from those used to safeguard any highly valuable or dangerous object: restricted access, armed guards, protective storage facilities, fences, motion detectors, and so on.

These safeguards are also applied throughout the personnel screening process. No individual is assumed to be completely reliable; hence, at least two persons are required to be present whenever access to a nuclear weapon is available.

On the other hand, preventing some individual, organization, or country from developing its own nuclear weapons is particularly difficult, especially in terms of

controlling critical technologies. Critical milestones for the production of nuclear warheads include:

- Production of fissile material through uranium enrichment programs
- Production of plutonium fissile material through a breeder reactor program or
- Procurement of existing bomb-grade materials and
- Design of high explosives to create powerful and precise implosions
- Development of neutron source technology
- Procurement of highly specialized technical knowledge required to produce high-yield nuclear weapons

Characteristics of Nuclear Explosions

A nuclear explosion, such as that depicted in Figure 19-6, is characterized by three main effects: a blast wave, a thermal pulse, and nuclear radiation. These three primary effects result in the most damage and, consequently, are the most relevant to predicting nuclear explosive effects. Secondary effects abound, but these effects are less useful in predicting the extent of overall damage.

Blast Wave

A thermonuclear bomb releases about 70 million times as much energy as the equivalent mass of conventional explosive. As discussed earlier in the text, a conventional explosion primarily releases its energy as heat, causing the combustion gases to expand and propel the casing fragments outward at high velocity while also compressing the surrounding air to create a shock wave. Nuclear explosions also release a large portion of their energy as heat. The difference between nuclear and conventional explosions is one of scale.

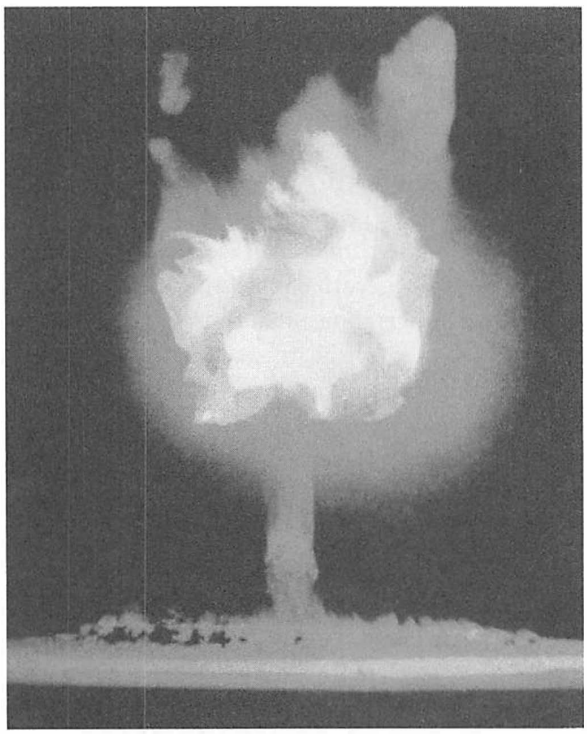


FIGURE 19-6 *Nuclear explosion*

Whereas a conventional explosion may reach several thousand degrees before the combustion gases begin to expand, a nuclear explosion may reach several millions of degrees before the gases expand. Recall that objects at temperatures of several thousand degrees will radiate energy at visible wavelengths. Recall also that the wavelength of maximum radiation can be found from Wein's Displacement Law:

$$\lambda = \frac{2{,}898\,\mu K}{T} \qquad (19\text{-}5)$$

where:

T = object temperature (in K)

If the temperature of a conventional explosion reaches 5,000°K, then the maximum radiation will be at 0.5796 nm, corresponding to yellow-green light. In contrast, a nuclear explosion where the temperature might reach 2 million degrees would produce a peak wavelength of 1.4 nm, corresponding with X rays.

Since X rays interact strongly with air, their energy is transferred to the surrounding air within a relatively short distance. As the surrounding air absorbs the X rays, it becomes very hot, and a fireball is created with diameters in excess of 1 mile. This superheated air in the fireball, between 3,000°C and 14,000°C, emits radiation with wavelengths in the infrared (IR) through the ultraviolet (UV) spectrum. This intense radiation, known as the thermal pulse, may be strong enough to spontaneously ignite flammable materials more than 10 miles away from the explosion center.

The rapid, outward expansion of the fireball creates a shock wave propagating from the point of detonation. Other than the strength and physical distance to which the shock wave effects will be felt, the shock wave from a nuclear explosion is no different from that produced by a conventional explosion.

From the scaling law introduced earlier, one can predict the distance to which a particular weapon's effects may be felt. For instance, for a 100-kiloton nuclear warhead (1 kT = 1,000 tons of TNT or 10^6 kilograms of TNT) compared with a 1,000-kilogram conventional bomb,

the effects extend approximately 10 times as far:

$$\left(\frac{W_1}{W_2}\right)^{1/3} = \left(\frac{1\times10^6\,\text{kg}}{1\times10^3\,\text{kg}}\right)^{1/3} = 10 \qquad (19\text{-}6)$$

All shock wave characteristics are magnified; the higher the peak overpressure, the more strongly the blast wind that follows. For conventional explosives, these winds might achieve 100 mph, but for nuclear explosions, maximum wind speeds can exceed 1,000 mph, and then slowly decay over the following 10 seconds.

Dynamic pressure from these winds can be devastating. For example, a 1,000 mph wind would create about 40 psi of dynamic pressure. This wind would create 56,000 pounds of force on a typical person, accelerating a 200-pound person at 10,000 ft/sec^2 (instant death). A 0.1 psi of dynamic pressure (corresponding to a wind speed of 70 mph) exerts about 150 pounds of force.

This wind would be felt at a range of almost 3 miles from a 100 kT explosion. Table 19-1 shows corresponding dynamic pressures for associated wind speeds.

The blast winds also change direction as the shock wave passes; see figure 19-7. This wind reversal occurs as the rapidly and outwardly expanding compressed air creates a localized negative pressure differential (rarefaction) over the space through which the shock wave has passed.

This negative overpressure results in an in-rushing movement of air toward the negative pressure. The reversal can be particularly devastating against structures that have flexed in one direction to accommodate the initial shock wave and then are rapidly flexed in the opposite direction.

TABLE 19-1 *Dynamic pressure as a function of maximum wind speed*

Dynamic Pressure (psi)	Maximum Wind Speed (mph)
0.1	70
1.0	210
10	590
100	1,600

Finally, as the negative pressure phase draws back toward the point of explosion, a much smaller localized overpressure may once again result in wind reversal, albeit a much milder one as pressure equalizes.

Unique to nuclear explosions is the afterwind caused by the enormous fireball rising in the atmosphere. As it rises, approaching speeds up to 300 mph, the surrounding air will be drawn into the vacuum created by the expanding and rising fireball. These winds, although strong, do not cause a great deal of damage in comparison to the aforementioned blast effects. However, the suction beneath the fireball draws along dirt and debris, forming the characteristic mushroom cloud as the fireball rises and cools; see figure 19-8.

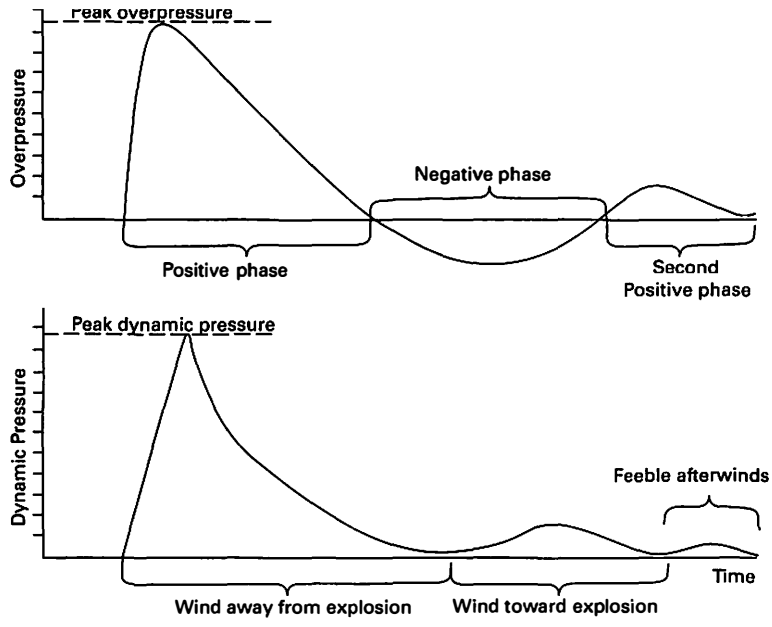


FIGURE 19-7 *Peak and dynamic overpressure as a shock wave passes*

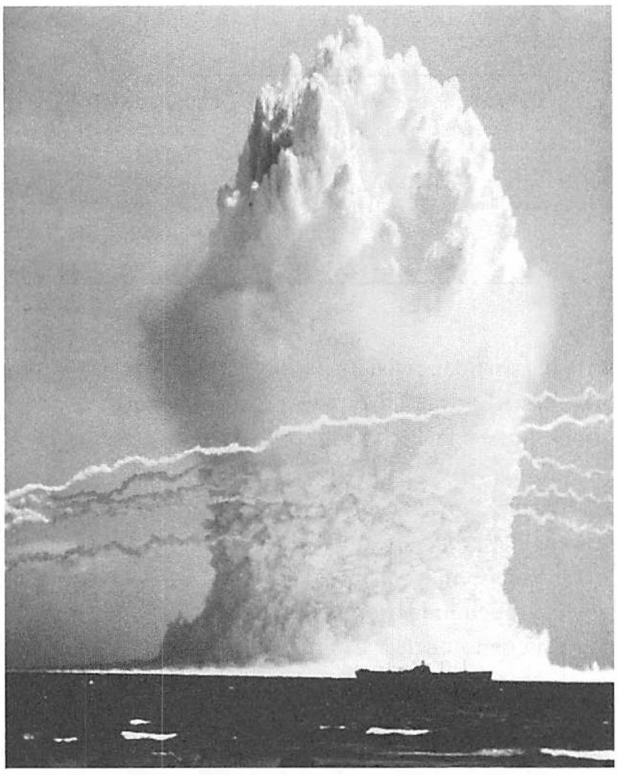


FIGURE 19-8 *Mushroom cloud beginning to form from a nuclear explosion in water*

Thermal Pulse

The fireball produced immediately after the detonation of a nuclear weapon will rapidly radiate energy in the form of thermal and visible light. The thermal energy effects will be quite strong, extending the farthest of the three main effects, and will be great enough to result in spontaneous ignition of combustible material.

Personnel exposed to a nuclear explosion will likely be burned. For a 100 kT warhead, personnel may experience first-degree burns up to 5 miles from ground zero (the point on the earth's surface that is directly under the detonation point). Approximately 10 seconds is required before the fireball's surface cools sufficiently to stop being a significant source of thermal energy.

Nuclear Radiation

Nuclear radiation is released from any nuclear explosion, the greatest effects coming from two types: neutron and gamma radiation. These two types are particularly harmful to life because of their ability to penetrate through most materials. Neutron radiation comes from the nuclear reaction itself as neutrons are produced as a by-product of fission reactions. Gamma radiation comes from three sources: directly from the nuclear reactions, radioactive decay of the fission by-products, and secondary decay reactions with materials that have absorbed neutrons.

The radiation from a nuclear explosion that is emitted within the first minute of the explosion, called prompt or initial radiation, is a combination of neutron and gamma rays. Residual radiation that is emitted after the first minute comes primarily from extensive gamma radiation from the radioactive fission by-products, which have relatively short half-lives.

Compared to thermal effects, the actual range to which the lethal radiation dose extends is relatively short. For the 100-kT warhead, personnel within 1.2 miles will receive a radiation dose that will likely cause death sometime in the next 30 days. These same people will also experience 10 psi peak overpressure and 300 mph winds as well as about 100 cal/cm^2 of thermal energy (10 cal/cm^2 is sufficient to cause third-degree burns). Any one of these effects in and of itself might be lethal.

Hence, the initial nuclear radiation effects are rarely the sole cause of death or destruction because some other lethal effect is also present at that range. Of major concern, however, is the longer-lived radioactive by-products that will contaminate the surrounding area. This radioactive contamination may be widespread because of the dispersion within the fireball and mushroom cloud. If it reaches upper-level winds, it may be carried far from the initial point of detonation.

Particularly long-lived isotopes can be absorbed by living organisms and cause a long-term risk from exposure. This contamination is impossible to detect without instrumentation and therefore will gradually be dispersed throughout water and food supplies as well as potentially be absorbed into the natural food chain.

The immediate vicinity of a nuclear explosion will remain radioactive for dozens of years. Additional physiological effects are discussed later in this chapter.

Additional Effects

The sheer magnitude of the energy released in a nuclear detonation creates a variety of secondary effects. The great flash of visible light immediately after the explosion may permanently or temporarily damage the retinas of personnel looking toward the detonation, depending on their distance from the blast.

The great electric potential developed by the uneven flow of ions outward from the explosion creates an electromagnetic pulse (EMP). The EMP has two primary effects: TREE and blackout. TREE is an acronym for Transient Radiation Effects on Electronics and is a result

TABLE 19-2 *Typical lethal ranges to exposed personnel for a 100 kT nuclear explosion*

Effect	Range (miles)
Blast overpressure (40 psi)	0.5
Blast wind (600 mph)	0.8
Radiation (500 rem)	1.4
Thermal pulse (10 cal/cm²)	3.1

Note: *rem stands for Roentgen equivalent to man and discussed with the measurement of radiation section.*

from induced voltages/currents caused by the pulse. The effect can temporarily or permanently damage electronics by inducing electronic voltages of up to 10,000 volts through the circuitry.

The EMP also causes a phenomenon known as blackout, which is caused by the ionization of the atmosphere, resulting in a temporary—several hours or days—interruption to radio communication.

The extent and magnitude of the EMP varies greatly; however, the greater the altitude of detonation, the more pronounced the EMP effects since the thinner atmosphere higher up results in the ionization of a greater portion of the atmosphere. Table 19-2 provides typical lethal ranges for various effects from a 100 kT warhead.

Types of Bursts

The relative effects of blast, heat, and nuclear radiation are largely determined by weapon detonation altitude. Nuclear explosions are classified as high altitude, air, surface, underwater, or subsurface bursts.

High-altitude burst A high-altitude burst occurs when the weapon is exploded at an altitude above 30 kilometers such that the initial X rays generated by the detonation dissipate energy as heat in a much larger volume of air molecules. Therefore the fireball is much larger and expands much more rapidly, but does not touch the ground.

The ionizing radiation from the high-altitude burst travels for hundreds of miles before being absorbed. Significant ionization of the ionosphere occurs, causing severe disruption in communications.

This burst also leads to the generation of an intense electromagnetic pulse, significantly degrading performance of or destroying sophisticated electronic equipment.

Air burst The weapon is detonated in the air at an altitude below 30 kilometers but at sufficient height that the fireball does not come in contact with the surface of the earth. After such a burst, the blast front may cause considerable damage and injury. The altitude of an air burst is often varied to achieve maximum blast effects by using the mach front, maximum thermal effects, desired radiation effects, or a balanced combination of these for a given yield warhead.

Burns to exposed skin may be produced over many square miles and eye injuries over a still larger area. Initial nuclear radiation is a significant hazard even with smaller weapons, but the fallout hazard can be ignored because there is essentially no local fallout from an air burst.

The fission products are generally dispersed over a large area of the globe unless there is local rainfall, resulting in localized fallout. In the vicinity of ground zero, there may be a small area of neutron-induced activity

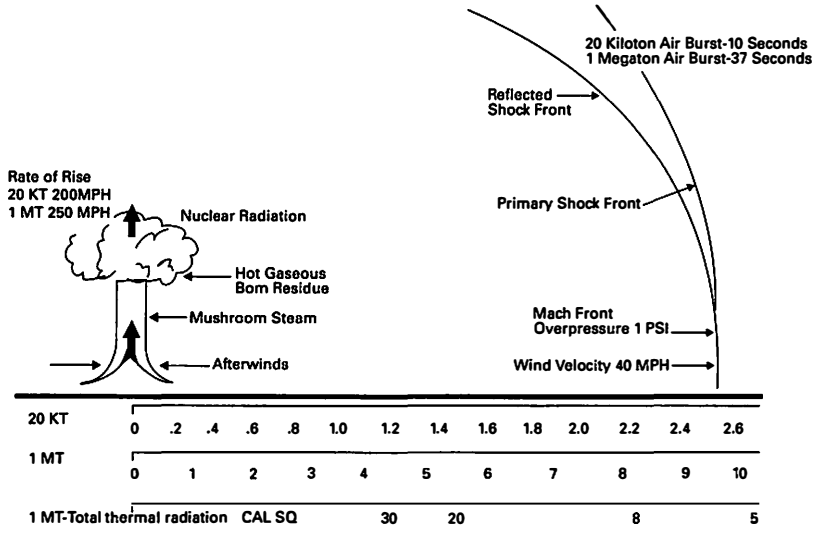


FIGURE 19-9 *Twenty kT airburst 10 seconds after detonation*

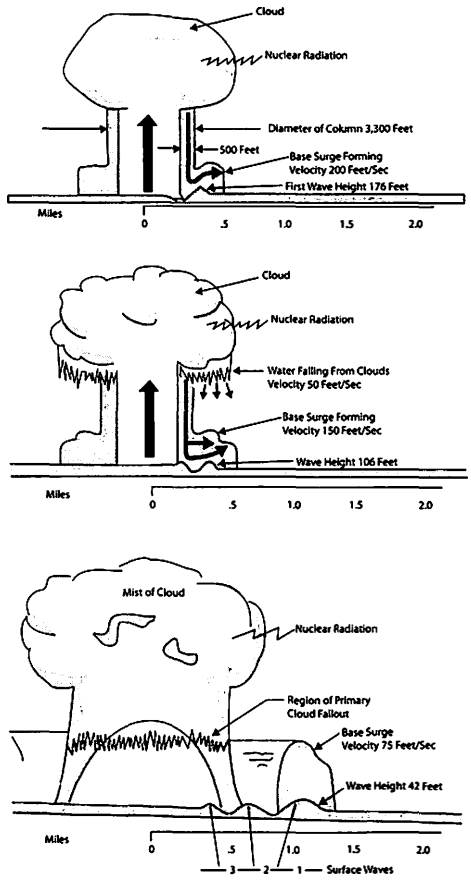


FIGURE 19-10 *Chronological development of a 100 kT shallow underwater burst*

that could be hazardous to troops required to pass through the area. Tactically, air bursts are the most likely to be used against ground forces.

Surface burst The weapon is detonated on or slightly above the surface of the earth so that the fireball actually touches the land or water surface. Under these conditions, the area affected by the blast, thermal radiation, and initial nuclear radiation will be less extensive than that for an air burst of similar yield since the earth's surface absorbs a significant portion of the energy from the detonation, except in the region of ground zero where this destruction is concentrated.

In contrast to air bursts, local fallout from a surface burst can be a hazard over a much larger downwind area than that which is affected by blast and thermal radiation since earth or water will be carried skyward due to the detonation's proximity to the surface and will consequently trap radioactive particles, which will then settle back toward the earth much more quickly than from an air burst. EMP effects are also very intense locally.

Surface bursts are generally used for single targets requiring maximum effect to provide requisite damage.

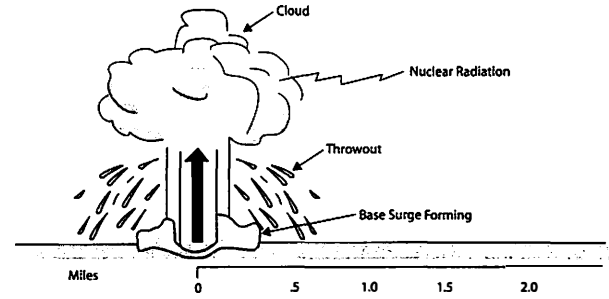


FIGURE 19-11 *100 kT shallow underground burst 45 seconds after detonation*

Underwater burst Underwater bursts create shock waves of greater peak overpressure but of shorter duration due to the denser medium in which the weapon is detonated. The peak overpressure is limited near the surface by destructive interference between the direct and reflected pressure waves from the surface.

Very deep detonations are used only for damaging submarines, whereas shallow underwater detonations are likely to damage surface ships. For shallow underwater bursts, where the fireball breaks the surface, a large ocean wave is experienced on the order of 100 to 200 feet that radiates outward a few miles from "ground" zero; see figure 19-10.

Subsurface burst (underground) The weapon is detonated beneath the surface of land. The formation of craters generally results from an underground burst, just as for a surface burst; see figure 19-11. If the subsurface burst does not break the surface, the only other hazard will be from ground shock. If the burst is shallow enough to break the surface, blast, thermal, and initial nuclear radiation effects will be present, but will be less than for a surface burst of comparable yield. However, local fallout will be very heavy in this situation.

Underground bursts are particularly useful against bunkered targets but require a penetrating warhead design.

Effects of Ionizing Radiation

When a uranium atom is split or fissioned, it does not always split in the same place every time. The two fission fragments are elements of lower atomic weight than uranium and contain the nucleons and electrons of the original larger uranium atom. The uranium atom, of course, ceases to exist after it is split, and more than 80 different possible fission products are formed. Each has the chemical properties associated with their new structure and the additional capability of being able to release ionizing radiation.

Ionization is any process that results in the removal of an electron from an atom or molecule, thereby leaving the atom or molecule, with a net positive charge. Ionization occurs

when sufficient energy is transferred to dislodge one of the electrons from the outer orbital shells of the target atom.

Ionizing radiation interacts with matter in one of two ways: it is either scattered or absorbed. Both result in deposition of energy in the target material. The mechanisms of absorption are of particular interest because absorption into body tissue may result in physiological injury, and the degree of absorption or type of interaction is a primary factor in determining shielding (protection) requirements.

With each fission, about two neutrons are released that can strike another fissile atom, causing more fissioning in what is usually called a chain reaction. Fission yields radioactive fission products, which emit radiation that causes existing chemicals in air, water, or other nearby materials to absorb energy, and changes their chemical structure slightly. Neutrons may also be absorbed in nearby materials, making them radioactive and affecting the chemistry of nearby materials. Approximately 300 different radioactive chemicals may be created with each chain reaction.

Besides their ability to give off ionizing radiation, many of the radioactive particles are biologically toxic for other reasons. Plutonium is chemically attracted to bone, as is the naturally occurring radioactive chemical radium. However, plutonium clumps on the surface of bone, delivering a concentrated dose of alpha radiation to surrounding cells, whereas radium diffuses homogeneously in bone and thus has a lesser localized cell damage effect. This makes plutonium much more biologically toxic than a comparable amount of radium.

The cellular damage caused by internally deposited radioactive particles becomes a health concern for the organ containing them. For example, radionuclides lodged in bones can damage bone marrow and cause bone cancers or leukemia, whereas radionuclides lodged in lungs can cause respiratory diseases.

Radiation Effects

Nuclear reactions will produce radioactive fission products as a by-product of the process. Four main types of radiation are produced from nuclear reactions and are capable of ionizing material that they pass through: gamma (or X-ray) radiation, neutrons, alpha particles, and beta particles.

Initial radiation, released within the first minute of a nuclear explosion, comes from the gamma rays and neutrons produced, the radioactive decay of short-lived fission products, and the interaction with the surrounding air from decays of isotopes that became radioactive from neutron absorption.

Residual radiation, which is produced after the first minute of a nuclear explosion, comes primarily from beta-gamma decay of the various longer-lived radioactive fission products and may be present for a significant period of time after detonation.

Gamma radiation These high-energy photons are actually short-wavelength X rays and may cause ionization via three main processes: photoelectric effect, Compton scattering, and pair production. In the photoelectric effect, the gamma photon interacts with and transfers all of its energy to an orbiting electron, ejecting that electron from the atom. The kinetic energy of the resulting photoelectron is equal to the energy of the incident gamma photon minus the binding energy of the electron.

Compton scattering is an interaction in which the incident gamma photon loses enough energy to an orbital electron to cause its ejection, with the remainder of the original photon's energy being emitted as a new, lower-energy gamma photon with an emission direction different from that of the incident gamma photon.

Pair production results when the gamma photon interacts in the vicinity of the Coulomb force of the nucleus and the energy of the incident photon is spontaneously converted into the mass of an electron-positron pair; a positron is a positively charged electron. Regardless of the method of ionization, the deposition of energy into the material with which it interacts causes serious biological damage to living cells. Whether an organism dies is directly dependent upon the amount of exposure to this radiation.

Gamma rays come directly from the fission reaction and account for approximately 3.5% of released energy. Gamma radiation is stopped most effectively by lead. Unfortunately, approximately 2 inches of lead is required to reduce radiation exposure by a factor of 10. This "tenth-thickness" is a relative measure of shielding necessary for various materials. For instance, for gamma radiation, 12 inches of concrete or 24 inches of water also equate to a "tenth-thickness" and could be used as substitute shielding for the 2 inches of lead.

Neutron radiation Although most ionizing radiation injuries associated with nuclear warfare are attributable to gamma radiation as previously described, a sufficient number of high-energy fission neutrons escape from the detonation to represent a significant hazard at considerable ranges.

The neutron is a particle and thus is fundamentally different from electromagnetic radiation. It also differs from other particulate radiations (alpha and beta) in that neutrons

do not carry any electrical charge. As a result, neutrons do not interact with the orbital electrons of atoms; instead, they interact directly with the nuclei of atoms, particularly those having low atomic mass numbers, by transferring their energy to target atoms and molecules through elastic and inelastic collisions with nuclei.

Since considerably more energy is transferred in a collision between two objects of similar mass, neutrons will transfer more energy to target nuclei of substances of low atomic mass, particularly hydrogen, rather than to heavy nuclei. Because of this, neutron radiation is very damaging to the human body, since 70% of the body is water. In some of these collisions, a proton will be ejected from the nucleus, which will then cause secondary ionization through electromagnetic interactions. Neutrons are released primarily as a direct product of fission and fusion reactions and not as residual radiation.

Water or other materials with a preponderance of light nuclei such as polypropylene shielding are most effective at stopping neutron radiation. About 10 inches of water or 24 inches of concrete equate to one tenth-thickness.

Alpha particles An alpha particle is a helium nucleus that consists of two protons and two neutrons strongly bound together by nuclear forces. Since they have no electrons, alpha particles are positively charged; they also have a mass of about 7,000 times that of electrons and are ejected from the nuclei of radioactive atoms.

Their strong positive charge causes ionization through electromagnetic interaction with other electrons and nuclei as they pass through material. Alphas interact strongly with all materials but penetrate only slightly since they quickly transfer their energy through these strong interactions.

Most alpha radiation comes from unused fission materials such as uranium and plutonium, which are naturally radioactive and decay through alpha particle emission. Alpha particles are stopped by the human epidermis (skin layer), but may damage it in the process.

Beta particles Beta particle decay involves the conversion of a neutron into a proton and an electron within the nucleus. Although the proton is retained in the nucleus, the beta particle (electron) is ejected with a velocity dependent upon its kinetic energy. Because of its smaller mass and relatively higher emission energies, a beta particle is less ionizing than an alpha particle but more penetrating.

Their electrical charge causes ionization through electromagnetic interaction with other electrons and nuclei, and they strongly interact with all types of materials. Most radioactive fission products decay by emitting beta particles and are the major source of beta radiation. Many radioactive elements are found in fallout and cause widespread radioactive contamination.

Beta particles are stopped effectively by almost any material, including clothing. The main hazard with beta radiation, as with alpha radiation, is internal exposure from ingestion or inhalation of contaminated materials containing radioactive particles that eventually decay by emitting a beta particle.

Both alpha and beta particles penetrate cell membranes more easily than they penetrate skin. Hence ingesting, inhaling, or absorbing radioactive chemicals capable of emitting alpha or beta particles—and thereby exposing delicate body parts such as the lungs, heart, brain, or kidneys to these particles—poses serious threats to human health.

Measuring Radiation

One method of radiation measurement is counting the number of nuclear disintegrations or decays that occur in a given unit of radioactive substance per second. This measure is usually standardized to the element radium, which was the first radioactive substance to be discovered and widely used. One gram of radium undergoes 3.7×10^{10} nuclear decays per second.

The activity of 1 gram of radium is called 1 curie (Ci), named for Madame Marie Curie, a Polish-born French chemist (1867–1934). In recent radiation protection guides, the curie is being replaced by the becquerel, which is defined as one atomic event per second. So, by definition, 1 gram of radium equals 1 curie or 3.7×10^{10} becquerels.

For example, it would take more than 1 million grams of uranium to have the equivalent radioactivity of 1 gram of radium. Both 1 million grams of uranium and 1 gram of radium would be measured as 1 Ci.

The energy released in nuclear decay is measured in several ways. The erg is a very small unit of energy. Lifting 1 gram of material 1 centimeter requires about 1,000 ergs of work. A material exposed to 100 ergs/g is said to absorb 1 *r*adiation *a*bsorbed *d*ose (rad). There is no direct conversion from curies, which are related to the number of atomic events, to the rad, which is energy absorbed in tissue.

Sometimes radioactivity is measured in counts per minute on a Geiger counter. A nuclear transformation within an energy range measured by the instrument and close enough to the instrument causes a count to be registered. It is important to note that most Geiger counters cannot detect alpha particle emitters like plutonium.

Units of Measure

In order to make meaningful statements about radiation, a system of units is needed. There are three main measures of radiation: the roentgen, the rad, and the rem.

Roentgen This is a measure of exposure to ionizing radiation. One roentgen is the quantity of gamma radiation that will cause 2.08×10^9 ion pairs to form in 1 cm^3 of dry air at standard conditions (STP). The roentgen is by definition applicable only to gamma rays or X rays.

Rad (radiation absorbed dose) This is a measure of dose, or how much radiation energy is absorbed. One rad represents the deposition of 100 ergs of radiation energy into 1 gram of material. In tissue at or near the surface of the body, gamma or X-ray exposure of 1 roentgen results in an absorbed dose of approximately 1 rad; but this rough equivalence does not apply to other materials.

$$1 \text{ roentgen exposure} = \begin{matrix}1 \text{ rad of dose in} \\ \text{body tissue}\end{matrix} \qquad (19\text{-}7)$$

Rem (roentgen equivalent to man) The real concern is always the radiation dose in human body tissue. For this reason, the most appropriate unit of dose is the rem, for "roentgen equivalent in man." Since the roentgen was defined for gamma radiation, a way to equate other types of radiation is needed. The rem dose is the rad dose times a quality factor Q, as shown in table 19-3. For immediate effects of radiation (radiation effects shortly after detonation), Q is usually taken to be 1, and rads and rems are used interchangeably.

To calculate the long-term biological damage (e.g., cataracts, leukemia, genetic changes) caused by radiation, including alpha emitters inside the body, the rad dose must be multiplied by Q to obtain the rem dose. An internally contained alpha particle does damage 20 times greater when lodged within a tissue, bone, or organ. For example, alpha particles giving a 2 rem (or rad) dose to skin would give a 40 rem dose to sensitive lung tissue when inhaled.

TABLE 19-3 *Quality factor (Q) for various radiation types*

Radiation Type	Q
Gamma	1
Thermal neutron	2
Fast neutron	10
Alpha (internal)	20
Beta	1

Theoretically, the rem dose measures equivalent biological effect, so damage from X rays, for example, would be the same as damage from alpha particles, when the dose in rem was the same. This is then used to derive the dose in human body tissue as follows:

$$\text{rem} = \text{rad} \times Q \qquad (19\text{-}8)$$

The rad and the rem are being replaced internationally by a pair of new units, the gray (Gy) and the sievert (Sv). The unit of absorbed dose, the gray, corresponds to 1 Joule of energy deposited per kilogram of tissue. Thus 1 Gy equals 100 rad and 1 centigray (cGy) equals 1 rad. The sievert is the corresponding unit of dose-equivalence and 1 Sv equals 100 rem and hence 1 centisievert (cSv) = 1 rem.

In order to have a quantitative sense of the frequency of the different cell effects caused by radiation exposure, imagine a colony of 1,000 living cells exposed to a 1 rad X ray, about the dose for one X-ray spinal examination. Two or three cell deaths, two or three mutations or irreparable changes in cell DNA, and about 100,000 ionizations in the whole colony of cells, ranging from 11 to 460 ionizations per cell, would occur. Cells can repair some, but not all, damage due to radiation.

A comparable 1 rad exposure to neutrons, which have a higher-quality factor, would be expected to cause more cell deaths and more mutations. The ionizations caused would range from 145 to 1,100 per cell. Alpha particles, which occur naturally, would cause roughly ten times as many cell deaths and mutations, and 3,700 to 4,500 ionizations per cell.

Somatic Effects

Ionizing radiation, by definition, creates ions as it passes through material. In human body tissue, the ions damage or alter cells in a variety of ways, including breaking chromosomes, causing the nucleus to swell, increasing the viscosity of the cytoplasm, increasing the permeability of the plasma membrane, and delaying or preventing mitosis.

Depending on which cells are damaged, a variety of effects may be produced on the entire organism. For whole-body exposure, certain organs and tissues are the most sensitive to exposure to ionizing radiation. For moderate exposures (100–1,000 rem), the blood-forming organs (hematopoietic tissue) are at greatest risk.

Moderate exposures to whole-body ionizing radiation pose great risk to the entire body by depleting the number of white blood cells, referred to as leukopenia, thus putting the body at risk to fight infection. If the radiation dose is high, extensive cell damage and health effects are seen immediately. Penetrating radiation doses

TABLE 19-4 *Probable health effects resulting from exposure to ionizing radiation*

Dose in Rems (whole body)	Health Effects Immediate	Delayed
1,000 or more	Immediate incapacitation and death.	None
600–1,000	Weakness, nausea, vomiting, and diarrhea followed by apparent improvement. After several days: fever, diarrhea, blood discharge from the bowels, hemorrhage of the larynx, trachea, bronchi, or lungs, vomiting of blood and blood in the urine.	Death in about 10 days. Autopsy shows destruction of hematopoietic tissues, including bone marrow, lymph nodes, and spleen; swelling and degeneration of epithelial cells of the intestines, genital organs, and endocrine glands.
250–600	Nausea, vomiting, diarrhea, epilation (loss of hair), weakness, malaise, vomiting of blood, bloody discharge from the bowels or kidneys, nose bleeding, bleeding from gums and genitals, subcutaneous bleeding, fever, inflammation of the pharynx and stomach, and menstrual abnormalities. Marked destruction of bone marrow, lymph nodes, and spleen causes decrease in blood cells especially granulocytes and thrombocytes.	Radiation-induced atrophy of the endocrine glands including the pituitary, thyroid, and adrenal glands. From the third to fifth week after exposure, death is closely correlated with degree of leukocytopenia. More than 50% die in this time period. Survivors experience keloids, ophthalmological disorders, blood dyscrasis, malignant tumors, and psychoneurological disturbances.
150–250	Nausea and vomiting on the first day. Diarrhea and probable skin burns. Apparent improvement for about 2 weeks thereafter. Fetal or embryonic death if pregnant.	Symptoms of malaise as indicated above. Persons in poor health prior to exposure, or those who develop a serious infection, may not survive. The healthy adult recovers to somewhat normal health in about 3 months. He or she may have permanent health damage, may develop cancer or benign tumors, and will probably have a shortened life span. Genetic and teratogenic effects.
50–150	Acute radiation sickness and burns are less severe than at the higher exposure dose. Spontaneous abortion or stillbirth.	Tissue damage effects are less severe. Reduction in lymphocytes and neutrophils leaves the individual temporarily very vulnerable to infection. There may be genetic damage to offspring, benign or malignant tumors, premature aging, and shortened life span. Genetic and teratogenic effects.
10–50	Most persons experience little or no immediate reaction. Sensitive individuals may experience radiation sickness.	Transient effects in lymphocytes and neutrophils. Premature aging, genetic effects, and some risk of tumors.
0–10	None	Premature aging, mild mutations in offspring, some risk of excess tumors. Genetic and teratogenic effects.

of 1,000 rad or more cause immediate brain death and paralysis of the central nervous system.

The radiation dose in which 50% lethality occurs within 30 days is referred to as LD 50/30 and equates to between 400 and 500 rem. Between 250 and 1,000 rad, death is usually due to damage to the stomach and gut. Below 250 rad, death is principally due to damage to the bone marrow and blood vessels. Penetrating radiation in doses above 100 rad inflicts severe skin burns. Vomiting and diarrhea are caused by doses above about 50 rad.

An individual may react differently at different times of life or under different circumstances. Below 30 rad, for most individuals, the effects from external penetrating radiation are not immediately felt. The mechanism of cell damage is similar to that described for minute quantities of radioactive chemicals that lodge within the body itself. Only when enough cells are damaged to interfere with the function of an organ or a body system does the effect become evident.

Subclinical dose (0–100 rem) For total dose levels of less than 100 rem, there is a greater than 90% survival rate. Although there may be some short-term effects, no medical treatment is required, so it is considered to be subclinical. At the high end of dose levels, near 100 rem, the victim may feel dizzy or nauseous, but will most likely recover within a few days.

Therapeutic dose (100–1,000 rem) For dose levels in the range of 100 to 1,000 rem, medical treatment improves the chance for survival, but death becomes increasingly probable with increased levels.

Palliative dose (1,000 rem or higher) For dose levels greater than 1,000 rem, the mortality rate exceeds 90%. For almost all cases, death is imminent, and treatment only moderates suffering. The greater the dose, the more quickly the effects will be felt. Above 5,000 rem, personnel will be immediately incapacitated.

All persons receiving a dose greater than 100 rem can be expected to suffer varying forms of radiation sickness. Radiation sickness has three distinct phases: initial, latent and final.

Initial phase During this phase, which begins anywhere from 15 minutes to 6 hours after the dose is received, the onset of symptoms begins. Typical symptoms include, nausea, vomiting, headache, dizziness, and general malaise. The initial phase lasts from 1 to 2 days.

Latent phase The latent phase, which lasts from 1 to 4 weeks, is marked by a temporary disappearance of all symptoms acquired in the initial phase. Often, the patient will feel fully recovered. However, most of the damage to the body will begin to take effect during this period. There is a reduction in the number of white blood cells (leukopenia), which will make the body susceptible to infection.

For exposures between 100 and 1,000 rem, it is possible for the individual to recover. The treatment must be designed to correct the damage to the hematopoietic tissue and fight off infection. Typical treatment options include blood transfusions, bone marrow transplants, and antibiotics.

Final phase The final phase, which lasts up to 8 weeks, is marked by a recurrence of all the symptoms acquired in the initial phase. Additional symptoms will appear including skin hemorrhages (purpurea), diarrhea, and hair loss (epilation). Since the body's immune system has been weakened, the most common cause of death is infection. The final phase can result in death, but complete recovery is possible. Table 19-4 summarizes the health effects associated with various exposure levels to ionizing radiation.

Protection Standards

Through the years, the upper limits on radiation limits have been set and continually revised by the International Commission on Radiation Protection (ICRP) and the National Council on Radiation Protection and Measurements (NCRP); see table 19-5. These standards are based upon conservative modeling of radiation effects, which has its origins with the study of survivors of the weapons dropped on Japan.

Radiation effects fall into two categories—deterministic and stochastic. Deterministic effects manifest themselves soon after exposure; these include skin burns, blood count effects, and cataracts. These effects are the result of severe dysfunction or death of the ionized cells within the tissue.

Stochastic effects are more subtle, due to the mutation of the DNA itself. Usually cancer is the long-term manifestation of these stochastic effects.

In order to place some perspective on these numbers, we examine some common sources of daily radiation. The cosmic radiation dose attributed to the sun accounts for about 25 mrem per year at sea level. This rate is doubled at the mile-high cities of Denver and Albuquerque and about quadrupled for Quito, Ecuador, which is at an elevation of 9,350 feet.

Gamma radiation from the decay of radioactive nuclides in the soil presents another 46 mrem per year on a worldwide average. Within the United States this number ranges from 15 to 150 mrem per year.

TABLE 19-5 *Various radiation limits set by the NCRP*

Current Standards Established by NCRP, 1993		
Category	**Annual Limit**	**Estimated Risk at the Annual Limit (cancer/genetic defects)**
Occupational annual whole-body limit for stochastic effects	5 rem (stochastic)	2 in 1,000 per year
Occupational lifetime limit	1 rem × age (years)	3 in 100 at age 70
Occupational annual limit for deterministic effects	15 rem to lens of eye 50 rem to any other organ or tissue	Minimal risk if limits not exceeded
Public annual whole-body limit for continuous exposure	100 mrem	1 in 10,000 per year
Public annual whole-body limit for infrequent exposure	500 mrem	1 in 10,000 per year
Negligible individual dose	1 mrem	5 in 10,000,000

Finally, indoor radon is the largest contributor on average to the annual background dose. The average home has been found to have about 1.25 picocuries per liter, which equates to an annual whole-body dose equivalent of 400 mrem (based upon staying consistently in the home, whereas it is 40 mrem based on 10% occupancy). Approximately 6% of U.S. homes exceed 4 picocuries per liter, which is the established level at which some corrective action should be taken.

Nuclear Weapon Damage Prediction

Damage Criterion

In order to successfully predict the range of damage for a nuclear explosion, the damage mechanism involved must first be known. Various targets are susceptible to different damage effects, and therefore this vulnerability can assist in deciding how best to attack these targets. Nuclear explosions create three main effects: a shock wave, a thermal pulse, and nuclear radiation. Targets may be vulnerable to one, two, or a combination of all of these effects.

Damage associated with the shock wave is divided into two categories, which were discussed under conventional weapon effects. Diffraction loading is associated with the peak overpressure of the blast wave, whereas drag loading and associated damage are attributed to the dynamic pressure effects. Although all structures are subjected to these two forces simultaneously within the blast wave, certain structures are more susceptible to one force over the other.

Diffraction loading The target is literally shattered by the effect of the incident blast wave; this shock wave envelops and diffracts around the target and the resultant force results in significant stress on structures. Furthermore, the target is not enveloped simultaneously on all sides, resulting in a differential pressure on each face of the structure as the wave engulfs the target. These differential pressures tend to crush objects since this high force is exerted on the structure over a particularly short duration and with sudden impact. The extent of damage depends on the peak overpressure. The damage criteria are identical to those for diffraction loading from conventional warheads, as described previously in chapter 18.

Drag loading Due to the high pressure associated with the shock front and the induced low pressure trailing the shock front, an intense wind is created. This wind imparts a lateral force on target structures, resulting in either translation or flexing of the structures by the dynamic pressure of the wind. Drag loading is dependent upon not only the dynamic pressure but also the physical dimension of the structure. Typical targets vulnerable to this effect are generally either lightweight or flexible.

The extent of damage depends on the peak dynamic pressure or the maximum wind speed. Again, the damage criteria are the same as for conventional warheads. Table 19-6 provides some correlation between the forces associated with peak overpressure, dynamic pressure, and wind velocity.

TABLE 19-6 *Overpressure, dynamic pressure, and wind velocity at sea level*

Peak Overpressure (psi)	Peak Dynamic Pressure (psi)	Maximum Wind Velocity (mph)
200	330	2,080
150	223	1,778
100	123	1,414
72	80	1,170
50	40	940
30	16	670
20	8	470
10	2	290
5	0.7	160
2	0.1	70

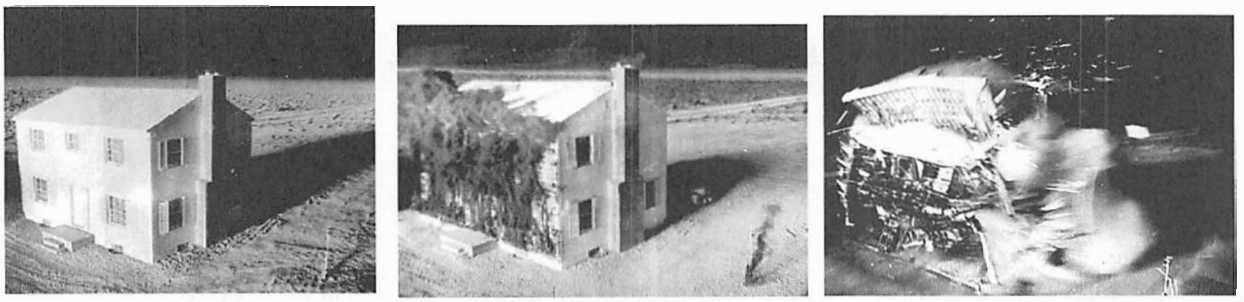


FIGURE 19-12 *This is a series of photos from a test. Notice thermal pulse arriving before blast front.*

TABLE 19-7 *Tabulated damage data for various targets*

Target Description	Damage Mechanism	Light Damage	Moderate Damage	Heavy Damage
Industrial buildings	Diffraction (peak overpressure)	3 psi	5 psi	15 psi
Roads and bridges	Diffraction (peak overpressure)	5 psi	8 psi	12 psi
Light armor	Drag (dynamic pressure)	1 psi	4 psi	7 psi
Heavy armor	Diffraction (peak overpressure)	10 psi	100 psi	200 psi
Personnel in open field	Drag (dynamic pressure)	1 psi	3 psi	5 psi
Personnel in bunkers	Diffraction (peak overpressure)	5 psi	30 psi	100 psi
Shallow, buried structures	Diffraction (peak overpressure)	30 psi	175 psi	300 psi
Parked aircraft	Drag (dynamic pressure)	0.7 psi	1.5 psi	3 psi
Ships	Drag (dynamic pressure)	2 psi	5 psi	7 psi
Residential structures	Fire Damage (thermal radiation)	10 cal/cm^2	20 cal/cm^2	50 cal/cm^2
Personnel	Burns (thermal radiation)	3 cal/cm^2	5 cal/cm^2	10 cal/cm^2
Personnel	Gamma/neutron radiation (nuclear radiation)	150 rem	500 rem	1000 rem

Fires/burns Nuclear bursts cause objects to ignite spontaneously or become charred. The extent of damage depends on the total thermal radiation exposure, measured in cal/cm^2. Wooden structures, fuel tanks, and personnel are particularly at risk from thermal radiation.

Table 19-7 provides typical values of thermal exposure necessary to cause light, moderate, or heavy damage. When specifically used to refer to personnel exposure, these terms correspond to first-, second-, and third-degree burns, respectively.

Radiation For purposes within this text, radiation exposure calculations apply only to the initial radiation dose (within the first minute of detonation). Only personnel are considered vulnerable to radiation exposure, although solid-state electronics may also be affected (referred to as transient radiation effects on electronics, or TREE).

Prediction Methodology The extent of damage depends on the total radiation dose received in the first minute. This is measured in units of rads or rem, and since only external exposure is assumed, the terms are used synonymously.

The first step in damage prediction is to identify the particular damage mechanism to which the target is susceptible. Next we determine the desired level of damage, which is subdivided into three categories for simplicity purposes: light, moderate, and heavy. By combining the information from the damage criteria tables with the necessary desired effect, the magnitude can be obtained for a given nuclear event.

Table 19-7 provides a summary listing and three columns of data representing several target types, the damage mechanism to which each is susceptible, and the necessary level of pressure, thermal, or radiation effects to achieve the respective level of damage.

Predicting Nuclear Blast Effects

The following four graphs, figures 19-13 through 19-16, depict various tabulated data representing shock wave effects, subdivided into curves showing peak overpressure and peak dynamic pressure, scaled for a 1-kiloton TNT equivalent explosion.

In order to apply these graphs to any warhead size, the scaling law presented earlier in the discussion of conventional weapons is applied. Since the reference warhead is 1 kT, the conventional equation is modified to:

$$\left(\frac{W_{\text{wrhd}}}{W_{\text{1kT}}}\right)^{1/3} = \left(\frac{h_{\text{wrhd}}}{h_{\text{1kT}}}\right) = \left(\frac{d_{\text{wrhd}}}{d_{\text{1kT}}}\right) \qquad (19\text{-}9)$$

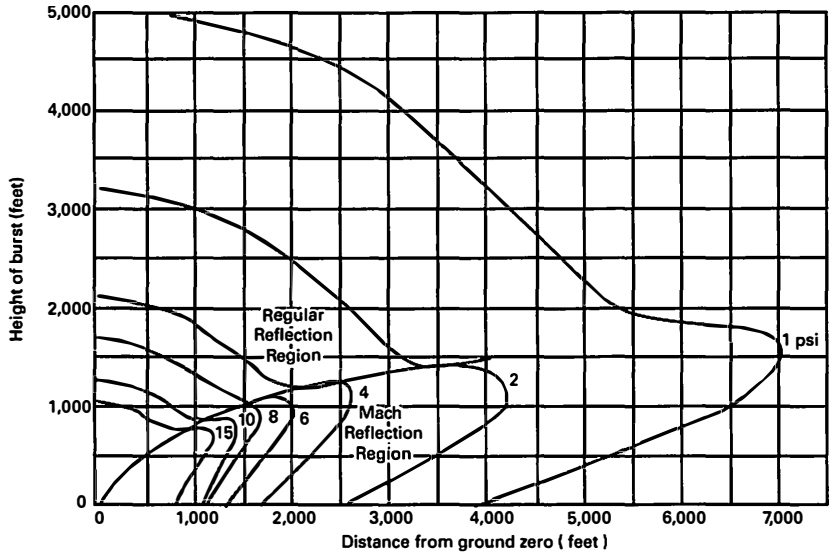


FIGURE 19-13 *Peak overpressure, low-pressure region*[i]

The numerators within this equation reflect the actual warhead size (W_{wrhd}), height of burst (h_{wrhd}), and the distance to which effects will be felt (d_{wrhd}) as measured from ground zero for the warhead of interest. The denominators represent height and distances as provided from respective graphs for the damage mechanism as referenced to a 1 kT burst.

EXAMPLE 19-1: Given a 1,000 kT warhead detonated 2,000 feet above the earth, find the peak overpressure 3,000 feet from ground zero.

SOLUTION: Applying the scaling law to determine the equivalent height of a 1 kT burst:

$$\left(\frac{W_{wrhd}}{1}\right)^{1/3} = \left(\frac{h_{wrhd}}{h_{1kT}}\right) \Rightarrow \left(\frac{1{,}000\,\text{kT}}{1}\right)^{1/3} = \left(\frac{2{,}000\,\text{ft}}{h_{1kT}}\right)$$

$$\Rightarrow h_{1kT} = \left(\frac{2{,}000\,\text{ft}}{(1{,}000\,\text{kT})^{1/3}}\right) = 200\,\text{ft}$$

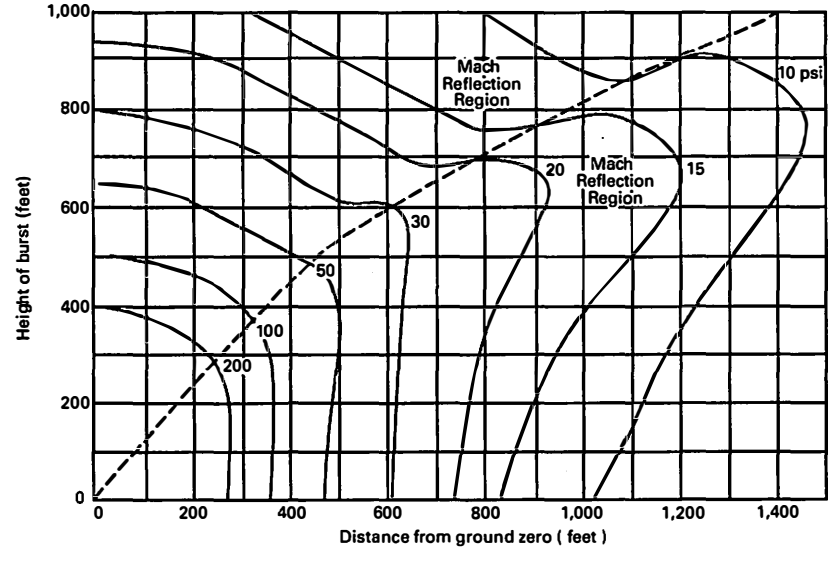


FIGURE 19-14 *Peak overpressure, intermediate-pressure region*[ii]

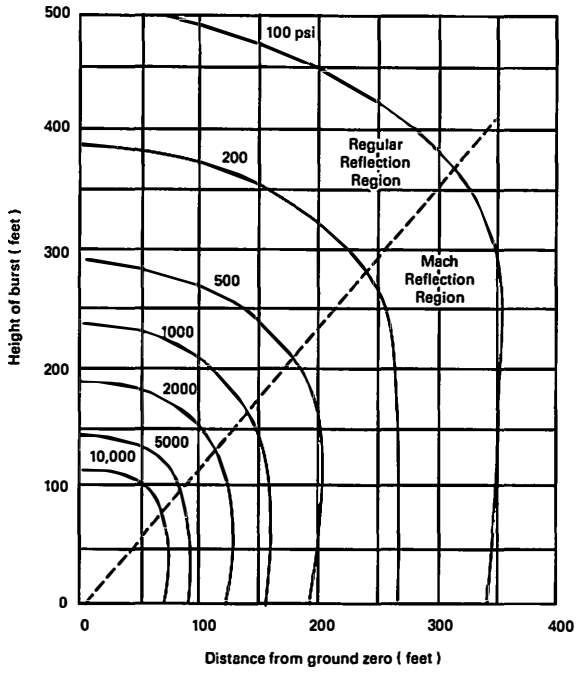


FIGURE 19-15 *Peak overpressure, high-pressure region*[iii]

Similarly, to determine the distance to which effects would be felt for a 1 kT burst:

$$\left(\frac{W_{\text{wrhd}}}{1}\right)^{1/3}=\left(\frac{d_{\text{wrhd}}}{d_{1\text{kT}}}\right)\Rightarrow\left(\frac{1{,}000\,\text{kT}}{1}\right)^{1/3}=\left(\frac{3{,}000\,\text{ft}}{\text{d}_{1\text{kT}}}\right)$$

$$\Rightarrow d_{1\text{kT}}=\left(\frac{3{,}000\,\text{ft}}{\left(1{,}000\,\text{kT}\right)^{1/3}}\right)=300\,\text{ft}$$

Effectively, this has converted the 1,000 kT warhead in the problem statement into equivalent heights and distances for a 1 kT warhead. Entering the appropriate curve using these parameters and noting where the intersection occurs on the respective graph provides the solution to the problem: 160 psi; see figure 19-17. Note that interpolation between the curves is required. For purposes of these calculations, assume that the spacing between the graph lines may be linearly interpolated.

EXAMPLE 19-2: At what optimum height of burst (HOB) should a 100 kT weapon be detonated such that the distance to cause heavy damage to industrial buildings is crossed? Optimum height of burst is defined as "that height on the pressure curves that maximizes distance to strike a target."

SOLUTION: Table 19-7 shows that "Industrial buildings" are damaged by diffraction loading (peak overpressure curves) and that 15 psi is necessary to cause heavy damage. Using figure 19-18 and the definition of optimum height of burst given above, the necessary information for a 1 kT warhead can be determined.

The optimum height of burst is approximately 660 feet for a 1 kT weapon. Applying the scaling law to scale up to the 100 kT weapon yields the following:

$$\left(\frac{W_{\text{wrhd}}}{1}\right)^{1/3}=\left(\frac{h_{\text{wrhd}}}{h_{1\text{kT}}}\right)\Rightarrow\left(\frac{100\,\text{kT}}{1}\right)^{1/3}=\left(\frac{h_{\text{wrhd}}}{660\,\text{ft}}\right)$$

$$\Rightarrow h_{100\text{kT}}=(660\,\text{ft})\times(100\,\text{kT})^{1/3}=3{,}063.4\,\text{ft}$$

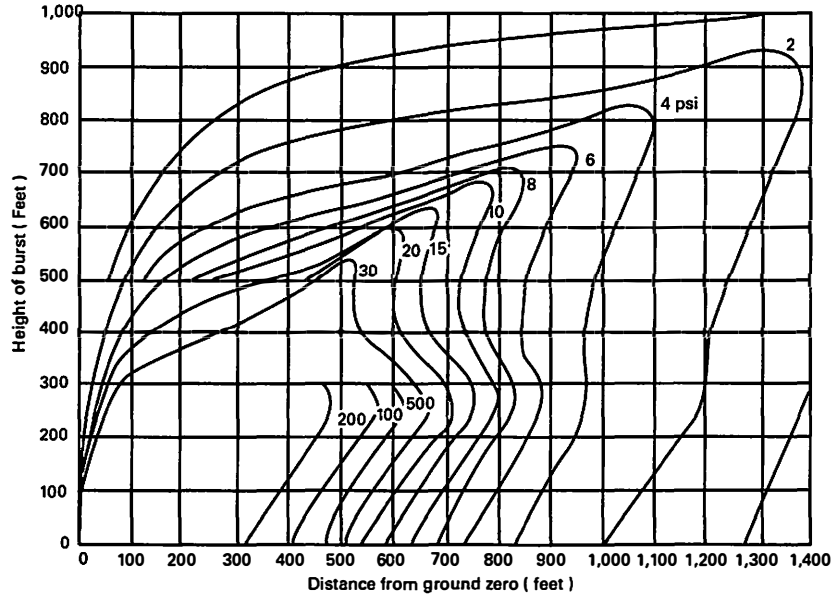


FIGURE 19-16 *Horizontal component of peak dynamic pressure for 1-kiloton burst*[iv]

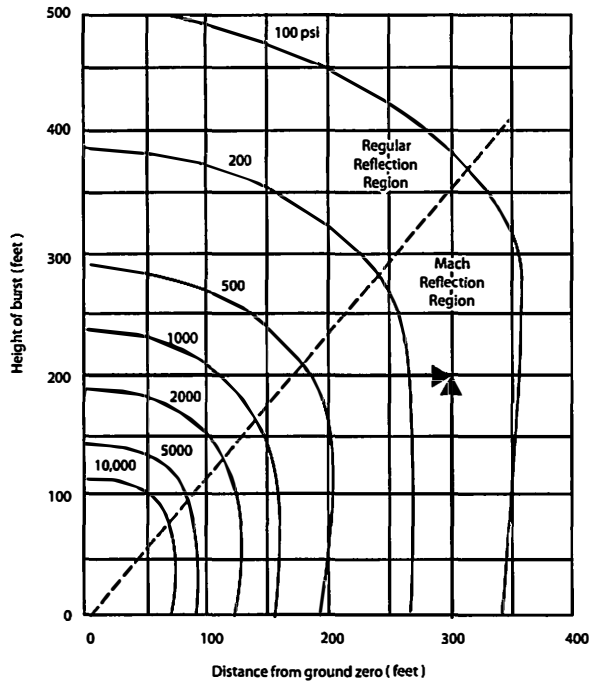


FIGURE 19-17 *Peak overpressure*

Thermal Radiation Prediction

Only one graph represents thermal effects requiring slant range and warhead yield as entering parameters. Finding slant range is nothing more than the application of the Pythagorean Theorem to the problem. Whereas the previous problems required distance measured from ground zero, slant range is actually the straight-line distance from detonation to the point where effects are being measured on the face of the earth.

Figure 19-19 represents the physical relationship between these concepts. Since warhead yield is included as an entering parameter, no application of the scaling law is needed.

Therefore, the slant range mathematical calculation is given as:

$$\text{slant range} = \sqrt{d^2 + (\text{HOB})^2} \qquad (19\text{-}10)$$

Note that slant range is expressed in different units depending on which particular graph is used; see figures 19-20, 19-21 and 19-22. For the thermal curves, slant range is expressed in miles. For neutron or gamma radiation, however, slant range is expressed in yards.

Figure 19-21 is displayed in miles because thermal effects extend much farther than blast or radiation effects. Note also that the curve doesn't go below 3 cal/cm^2 since thermal exposures less than this are biologically insignificant. Likewise, no curves extend beyond 50 cal/cm^2 because exposure to that level is sufficient to cause massive (lethal) damage to any vulnerable target.

EXAMPLE 19-3: Given a 75 kT warhead detonated at 300 feet, find the radiant (thermal) exposure 2 miles away.

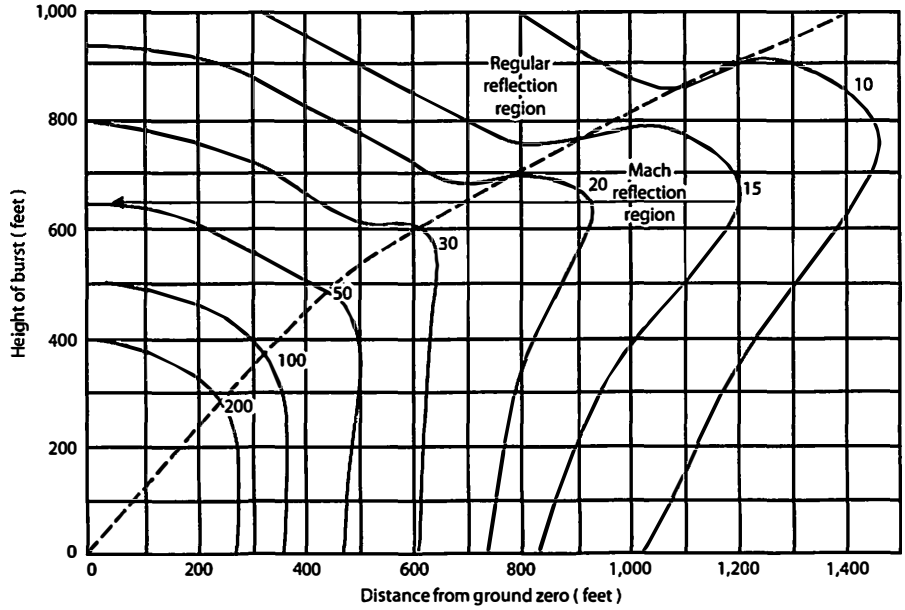


FIGURE 19-18 *Peak overpressure*

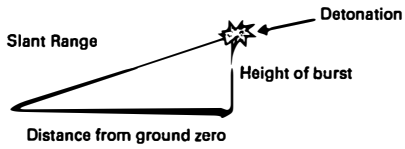


FIGURE 19-19 *The relationship between slant range and HOB*

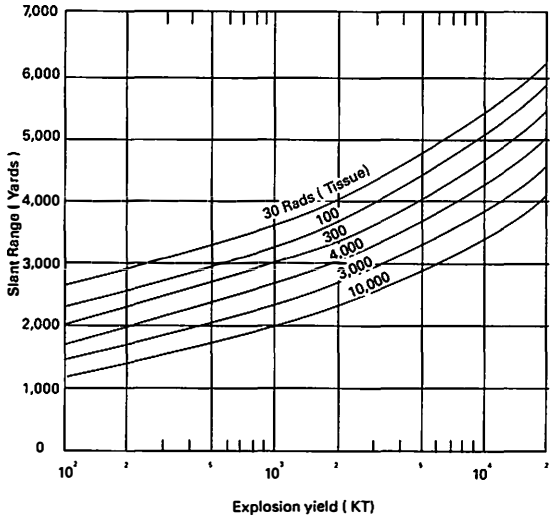


FIGURE 19-20 *Slant ranges for specified gamma-ray doses for targets near the ground as a function of energy yield of air-burst thermonuclear weapons with 50% fission yield, based upon 0.9 sea level air density*[v]

SOLUTION: Computing slant range from the aforementioned relationship:

$$\text{slant range} = \sqrt{d^2 + (\text{HOB})^2}$$

$$= \sqrt{(2 \text{ miles})^2 + \left(\frac{300 \text{ ft}}{5{,}280 \text{ ft/mile}}\right)^2}$$

$$= 2.00 \text{ miles}$$

Entering figure 19–21 with a slant range of 2.00 miles and a warhead yield of 100 kT provides the solution: about 22 cal/cm^2. Interpolation between lines may be necessary. Assume the interpolation between any two individual lines to be linear.

Radiation Dose Prediction

Multiple graphs are also available for ionizing radiation dose determination. The applications to these graphs are determined similarly to how thermal dose is determined.

Both figures 19-22 and 19-23 predict the dose in rads. The graphs don't go below 30 rads because dose levels lower than that will not cause any noticeable somatic effects. Likewise, for doses above 10,000 rads (rem), immediate death will occur in all cases.

EXAMPLE 19-4: Given a 200 kT warhead detonated at 750 feet, determine the gamma radiation dose at 2,000 yards from ground zero.

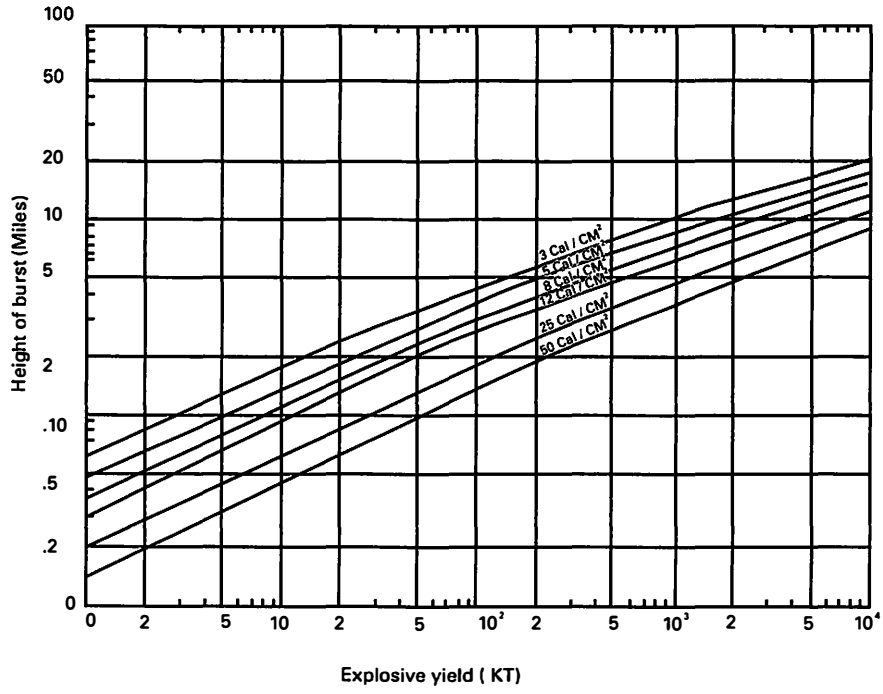


FIGURE 19-21 *Slant ranges for specified radiant thermal exposures on the ground as a function of energy yield of air burst at altitudes up to 15,000 feet for a 12-mile visibility*[vi]

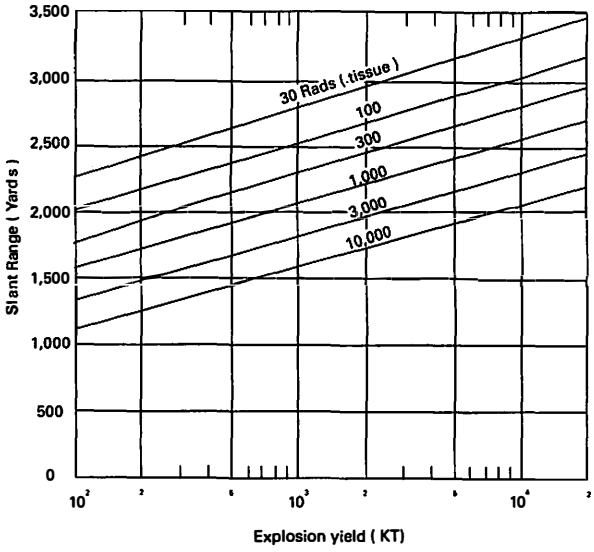


FIGURE 19-22 *Slant ranges for specified neutron doses for targets near the ground as a function of energy yield of air-burst thermonuclear weapons based on 0.9 sea level air density*[vii]

SOLUTION: Calculating slant range (on the graph of interest; see figure 19-23, slant range is measured in yards):

$$\text{slant range} = \sqrt{d^2 + (\text{HOB})^2}$$

$$= \sqrt{(2{,}000\ \text{yds})^2 + \left(\frac{750\ \text{ft}}{3\ \text{ft/yd}}\right)^2}$$

$$= 2{,}015.6\ \text{yd}$$

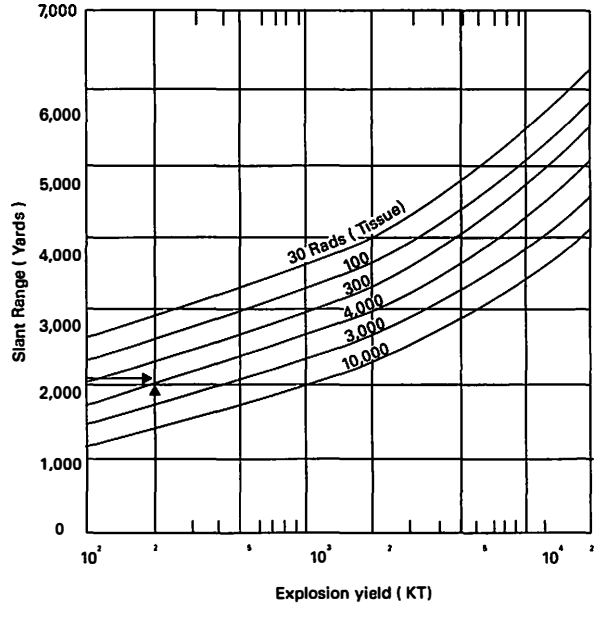


FIGURE 19-23 *Slant ranges for specified gamma-ray doses for targets near the ground as a function of energy yield of air-burst thermonuclear weapons with 50% fission yield, based upon 0.9 sea level air density*[viii]

Entering the figure 19-23 with a slant range of 2,015.6 yards and a warhead yield of 200 kT provides the appropriate answer: 925 rem.

Chemical and Biological Weapons (CBW)

The CBW threat is multidimensional in terms of the diversity of potential users and the circumstances of that use. CBW can be used by rival nation-states or terrorist organizations and can be employed on the battlefield or directed against a civilian population. Traditionally, the perceived threat of CBW was directed toward combat troops. Today, however, the use of CBW against domestic targets is becoming a more credible threat.

Terrorist use of chemical or biological weapons is among the most alarming of emerging transnational threats. Both the absence of other dominating global powers and the existence of overwhelming U.S. military capability greatly limit a terrorist's options. Increasing numbers of terrorist groups are looking to make use of asymmetric measures to accomplish their goals.

The international community adopted the Chemical Weapons Convention and the Biological Weapons Convention in an attempt to prevent the proliferation of chemical and biological weapons. Despite these efforts, treaty shortcomings continue to exacerbate the problems of proliferation.

The language of the treaties focuses on large developmental programs in sovereign states and never were intended to deal with the small quantities of chemical or biological agents that might be used by terrorists. Additionally, the treaties do not effectively regulate "dual-use" items, which are materials that can be used to manufacture chemical or biological weapons but that are also instrumental to basic production needs within the chemical or pharmaceutical industries.

Because of this, the chemical weapon convention does not address many toxic industrial chemicals that might be used as weapons. Additionally, the treaties were developed when these weapons were considered difficult to create or control. Advances in technology and genetic engineering have provided the ability to produce and manipulate these agents, making them effective tactical and operational weapons.

CBW Characteristics

Unlike nuclear or conventional weapons, chemical and biological weapons cause incapacitation or death solely through their interaction with the human body.

TABLE 19-8 *General comparison of chemical and biological agents*

Characteristics	Chemical	Biological
Area affected	Relatively small (tactical)	Small to very large (strategic), depending on spread of agent
Detectability	Difficult	Very difficult due to incubation period and necessity for symptoms to appear
Time to detect/identify	Seconds; several agent tests exist that provide rapid results	Tens of minutes; requires extensive medical facilities
Time until onset of effects	Normally minutes	Normally days
Medical treatment	Limited; may be able to treat only symptoms	Effective to nonexistent, depending on agent

Chemical weapons are poisons that affect the living system after introduction through the skin, eyes, lungs, blood, nerves, and other organs.

Biological weapons are either disease-causing microorganisms—such as bacteria, rickettsia, and viruses—or toxins, which are chemicals produced by living organisms that interact with humans, plants, or animals in the same way as chemical poisons.

Rudimentary chemical and biological weapons require minimal technology and are available to any state desiring to produce them. The most basic chemical warfare agent, mustard gas (levinstein mustard), was first synthesized in 1823. Advanced chemical or biological weapons, including high-purity nerve agents, are available to almost any country and organizational regime with chemical engineering, pharmaceutical, and biotechnology industries.

Access to chemical and biological technology is growing due to a variety of factors including economic conditions, the globalization of biotechnology and pharmaceutical industries, and scientific and engineering personnel exchanges that have facilitated the transfer of information. Table 19–8 provides a summary of the general characteristics of chemical and biological weapons.

The unique nature of chemical and biological weapons requires an integrated defensive posture to counter the threat. The threat from these weapons is unique, not in that they cause death or incapacitation but in how they cause it.

Chemical and biological weapons also invoke fear among the attacked party, especially if it is unprepared, because these weapons are often invisible, thus striking without warning. When airborne, such as in a smoke or vapor cloud, these weapons are amorphous and cannot be dodged like a bullet. They can literally seep through cracks and crevices.

Although chemical and biological weapons are in themselves designed to be deadly, it is the method and accuracy of their delivery that determine the severity of the damage. The ultimate effectiveness of these weapons is determined by the following factors, regardless of the circumstances of the weapons use:

- Agent delivery
- Doses on target
- Downwind dispersal
- Doses inhaled or absorbed
- Symptoms
- Performance degradation

Agent delivery Delivery methods of CBW are virtually unlimited. Battlefield delivery systems may include cruise or ballistic missiles, artillery, rockets, mortars, mines, bombs, or spray tanks. Defenses against agent delivery include missile defenses and preemptive strikes against CBW sites. Terrorist delivery systems are limited only by the imagination.

The dispersal of chemical and biological weapons can take a wide variety of forms. Probable means of weapon employment include:

- Off-target (upwind) attacks using agent aerosol disseminators moved along paths perpendicular to wind direction. Means of delivery could be aircraft, UAVs, cruise missiles, boats/submersibles, or ground vehicles. Such attacks could also be achieved with multiple-source detonation/spray devices covertly placed upwind from the target and triggered remotely or by timing devices.
- On-target attacks using various forms of agent containing fused munitions that explosively disseminate or spray the agent at or near ground level. Among these munitions are ballistic and cruise missile warheads, aircraft ordnance, tube and rocket artillery, and naval gunfire.
- Area-denial attack using persistent, generally chemical agents laid down in a heavy pattern with the

intention of contaminating ground areas and water-crossing points that enemy forces may attempt to traverse. Means of delivery include aircraft ordnance, artillery, and mines.

Doses on target Once delivered, it is critical to determine how much agent is present on target and in what form (vapor or liquid). From an adversary's perspective, this is the most critical and perhaps the most difficult factor in employing chemical or biological weapons.

This factor highlights the complexity of using chemical and biological agents as weapons. Some agents have extreme lethality, yet are difficult to aerosolize—for instance, botulinum toxin—whereas other agents have very low lethality but may be selected as a weapon because of the ease of aerosolizing them—for instance, tularemia. Defense includes early warning systems or tactics to allow avoidance to exposure.

Downwind dispersal Dispersal depends on the type of agent, the delivery system, and the geographical and meteorological conditions. A thickened vesicant agent on a cold, windless night poses little downwind hazard but poses the greatest threat at the point of impact or dispersion, whereas a biological agent such as anthrax, dispersed under optimum conditions from a single sprayer, can create a lethal downwind hazard area of more than a thousand square kilometers.

The nature of the delivery, therefore, may allow for the dispersal of an agent many kilometers from the intended target site, thus increasing the difficulty in preventing or detecting the agent's release. Detection and warning systems are needed to defend against this hazard. Decontamination systems may also be used to prevent the spread or reaerosolization of agents.

Doses inhaled or absorbed Once doses have been delivered, personnel are exposed to the hazards of chemical and biological agents by inhalation or percutaneously (through the skin). Biological agents primarily pose an inhalation hazard, although absorption is a possibility. Chemical agents may have lethal effects through both contact and inhalation. Defenses against this aspect of the threat include masks, suits, gloves, and collective protection systems and shelters.

Symptoms If personnel are exposed to chemical or biological agents, various symptoms may occur. Symptoms may be immediate—in the case of nerve agents—or delayed by several days—in the case of biological agents. Symptoms range from nausea and dizziness to paralysis and death. Vaccines, if they exist, may prevent or minimize the effects of exposure to chemical or biological agents. Therapies and drugs may be available for treatment when vaccines are not available for prevention.

Performance degradation If defenses are not effective and personnel develop symptoms from exposure, the performance of forces may be significantly degraded. Effective casualty management can reduce the number of casualties and expedite the return to duty. Training can reduce the performance degradation by teaching personnel to wear defensive equipment or not to develop a psychological fear of chemical or biological agents.

In a battlefield setting, several additional elements can become relevant. Future scenarios for the use of chemical and biological warfare against engaged forces are not expected to differ from those envisioned historically in conjunction with large-scale chemical and biological warfare programs that a number of countries have pursued.

Of the targets listed below, those in the first category tend to be most susceptible to infectious agents, which have a relatively slow onset of effect but a larger area of coverage. Targets in the second and third categories are vulnerable to a wide variety of chemical and biological agents. The targets in the fourth category are most susceptible to chemical- and toxin-type agents, which have a relatively rapid onset of effect but a smaller area of coverage per unit weight of agent.

1. High-value, large-area facilities and targets within or outside the theater operations: leadership, diplomatic, military headquarters; industrial, commercial, population centers
2. Theater-support military facilities: command and control, troop barracks, air bases, missile launch sites, naval ports, logistical transfer/storage facilities
3. Military assets near engagement areas: troop convoys, staging areas, drop zones, airstrips, air defense systems, artillery support bases, naval task forces
4. Engaged forces: infantry, amphibious, mechanized/armor

Aimed at certain critical nodes in a military's infrastructure, either domestically or abroad, chemical or biological weapons could disrupt the execution of military objectives. Therefore, it is imperative that military units have the ability to operate effectively in a contaminated environment while simultaneously being able to identify threat agent(s), treat injured personnel, and remediate the contaminated area.

Chemical Warfare Agents

A chemical agent is defined as "a chemical substance intended for use in military operations to kill, seriously injure, or incapacitate through its physiological effects." Chemical agents are essentially poisons affecting living systems through the skin, eyes, lungs, blood, nervous systems, and other organs.

These agents are often divided into four basic categories: nerve, blood, choking, and blister agents. Lethal agents, as the name implies, primarily cause death to exposed personnel and are generally from one of the first three categories.

Blister agents, however, primarily damage any tissue that they come in contact with and cause skin, eye, or respiratory damage, and may be employed to contaminate terrain, facilities, and material.

Chemical agents are classified as nonpersistent, as in the case of the G-series of nerve agents, or persistent, as with the nerve agent VX and blister agents. Persistency refers to the duration of an agent's effectiveness under the environmental conditions after dissemination and is typically measured by its duration in the battlefield environment once deployed.

Persistent agents generally present both a contact hazard and an inhalation hazard, whereas nonpersistent agents present only an inhalation hazard. This can be achieved by dissolving polymers—polystyrene or rubber products—in the chemical warfare agent to make the product highly viscous or thickened. The result is that the agent's persistence and adhesive ability increases, complicating decontamination. Table 19-9 provides a summary of various chemical warfare agents' properties.

Nerve Agents

Nerve agents disrupt the transmission of impulses within the nervous system by inhibiting enzymes at nerve junctions. Stable, easily dispersed, and highly toxic, nerve agents have rapid effects within the human body regardless of whether they are absorbed or inhaled. All nerve agents in the pure state are colorless liquids.

The ingredients (called precursors) of nerve agents, although controlled by international treaties such as the Chemical Weapons Convention, are easily obtainable. The most common nerve agents are: tabun (GA), sarin (GB), soman (GD), cyclosarin (GF), and VX.

Extremely toxic and rapidly effective, nerve agents enter the body through inhalation, absorption through the skin, or consumption of contaminated food or water. The time until the onset of nerve agent effects is a function of the method of exposure. Respiratory exposure is generally the most rapid debilatory method since the agent can rapidly diffuse into the blood circulation system.

Symptoms of nerve agent poisoning include increased production of saliva, chest pain, miosis (constriction of the pupil), headache, fatigue, slurred speech, hallucinations, nausea, difficulty breathing, coughing, cramping and vomiting, local tremors, convulsions, and loss of consciousness. The toxic effects depend on both the concentration of the nerve agent inhaled from the air and the time of exposure.

If medical treatment is to have any beneficial effect, it must be started immediately. Autoinjectors containing antidotes to nerve agents such as atropine are the most commonly encountered form of treatment.

TABLE 19-9 *Chemical warfare agents*

Class of Agent	Symbol	Symptoms	Effects	Rate of Action
Nerve	GA GB GD GF VX	Difficulty breathing; sweating, drooling, convulsions, dimming of vision	Incapacitates at low concentrations; kills in higher dosages. VX is persistent; others are considered nonpersistent.	Vapors: seconds to minutes Skin: 2 to 18 hours
Blood	AC CK	Rapid breathing, convulsions, coma	Kills in sufficient dosage. Nonpersistent.	Immediate
Blister	HD HN HL L	No early symptoms; searing/stinging of eyes/skin/mucosal membranes	Blisters (delayed hours to days); immediate pain. Persistent.	Vapors: 4 to 6 hours Skin: 2 to 48 hours
Choking	CG DP	Difficulty breathing; tearing of the eyes	Damages/floods lungs. Nonpersistent.	Immediate to 3 hours

Subsequently, the victim should be treated by qualified medical staff, who should initially inject additional antidotes and an anticonvulsant such as diazepam. Recovering from severe nerve agent poisoning requires at least 2 weeks and is characterized by amnesia, anxiety, muscular weakness, and difficulties in sleeping and concentrating.

Blood Agents

Most blood agents are nonpersistent and are primarily absorbed by breathing. These agents block the body's ability to absorb oxygen by affecting the enzyme in blood that is necessary for oxygen transport. Two primary blood agents exist: hydrogen cyanide (AC) and cyanogen chloride (CK).

Both gaseous and liquid hydrogen cyanide can also enter the body through skin absorption, although the agent's volatility makes open battlefield use problematic. On the other hand, hydrogen cyanide concentration may rapidly reach lethal levels if released in confined spaces.

Symptoms of cyanide poisoning vary and depend on route of poisoning, total dose, and exposure time. If hydrogen cyanide has been inhaled, the initial symptoms are restlessness and an increased respiratory rate. Other early symptoms are giddiness, headache, palpitations, and respiratory difficulty, followed by vomiting, convulsions, respiratory failure, and unconsciousness. If the poisoning occurs rapidly as a result of extremely high concentrations of cyanide in the air, symptoms may not be exhibited prior to death.

No medical antidote against cyanide poisoning exists. Current treatment is based on speeding up the body's own ability to excrete cyanide and to bind cyanide in the blood.

Choking Agents

Choking agents are designed to cause death in an exposed individual and injure unprotected personnel mainly through the respiratory tract (nose, throat, and lungs). These agents will "choke" an unprotected person. Upon exposure, membranes swell and secrete fluid, the lungs fill with this fluid, and death results from a lack of oxygen. The term for such a death is *dry land drowning.*

Two main choking agents exist: phosgene (CG) and diphosgene (DP). CG is a chemical agent with a short agent-cloud duration. Diphosgene (DP) is also a colorless gas with an odor similar to CG. DP has a stronger tearing effect than CG and thus has less of a surprise value when used against personnel. Its symptoms and effects are similar to those of CG.

Blister Agents

Blister agents, also called vesicants, cause inflammation, blisters, and general distress to tissue. Vulnerable areas include the eyes, respiratory system, and internal organs as well as exposed tissue. Blister agent effects are delayed and the first symptoms do not occur until 2 to 24 hours after exposure.

The severity of the blister is directly related to the agent concentration and the duration of contact. Since they are persistent in nature, blister agents can also be used to contaminate supplies or facilities. Blister agents are divided into three types: mustards, arsenicals, and urticants.

Mustards In its pure state, a mustard agent is colorless and almost odorless. The name *mustard* was given to blister agent "H" as a result of an earlier production method that yielded an impure, mustard-smelling product. Mustard agent is also claimed to have a characteristic smell similar to rotten onions. However, the sense of smell is dulled after only a few breaths, and the smell can no longer be distinguished. In addition, mustard agent can cause injury to the respiratory system in concentrations that are so low that the human sense of smell cannot detect them.

Whether in gas or liquid form, mustard agents attack the skin, eyes, lungs, and gastrointestinal tract. Internal organs may also be injured as a result of the agent being absorbed through the skin or lungs and transported into the body.

Delayed effects are a characteristic of mustard agents. Mustard agent give no immediate symptoms upon contact, and consequently a delay of between 2 and 24 hours may occur before pain is felt and the victim becomes aware of what has happened. By then, however, cell damage has already occurred.

Symptoms of mustard agent poisoning extend over a wide range. Mild injuries consist of aching eyes with or on abundant flow of tears, inflammation of the skin, irritation of the mucous membranes, hoarseness, coughing, and sneezing. Normally, these injuries do not require medical treatment. Severe injuries, which are incapacitating and require medical care, may involve eye injuries with loss of sight, the formation of blisters on the skin, nausea, vomiting, and diarrhea together with severe respiration difficulty.

No treatment or antidote exists to counteract the basic cause of mustard agent injury. Instead, symptoms must be treated. By far the most important measure is rapidly and thoroughly decontaminating the patient and thereby

preventing further exposure. This decontamination will also decrease the risk of exposure to medical staff.

In decontamination, clothes are removed; the skin is decontaminated with a suitable decontaminant and washed with soap and water; hair is shaved off if it is contaminated; and eyes are rinsed with water or a physiological salt solution for at least 5 minutes.

In medical treatment, efforts are made to control infections by using antibiotics. Pain can be eased with local anesthetics. After skin injuries have healed, it may be necessary to perform plastic surgery. Lung injuries are treated with bronchodilatory treatment. Medicine to relieve coughing and cortisone preparations may be used. Eye injuries are treated locally with painkillers and with antibiotics if required. Despite treatment, inflammation and sensitivity to light may remain for long periods.

Arsenicals The arsenicals are a group of blister agents that have arsenic as a central element of their chemical structure. Arsenicals produce much the same injuries to the skin and mucous membranes as mustards, but have the added effect of being a systemic poison. Arsenicals are colorless to brown liquids that evaporate more quickly than mustards and have a fruity or geraniumlike odor. They are much more dangerous as liquids than as vapors.

Immediate decontamination to remove the liquid agent is necessary in contaminated individuals, but it is not necessary for exposure to vapors unless pain is present. Sneezing and irritation to the upper respiratory tract can result from exposure to the vapors. There are three main arsenicals: Lewisite (L), mustard-lewisite mixture (HL), and phenyldichloroarisine (PD).

Urticants Urticants are blister agents that cause an immediate, severe burning sensation followed by intense pain and then a feeling of numbness. The most common urticant is phosgene oxime (CX). It has a disagreeable, penetrating odor and can appear as a colorless crystalline solid or as a liquid. CX causes violent irritation to the mucous membranes of the eyes and nose.

An individual exposed to CX will first develop an area of pale skin surrounded by a red ring where the agent came in contact with the skin. A welt resembling a bee sting will form within 30 minutes. The area will turn brown within 24 hours, and a scab will form within a week. Healing could be delayed for as long as 2 months. Any skin exposed to CX should be decontaminated as soon as possible by flushing the area with large amounts of water.

Agent Production

Every country with a chemical industry is capable of creating blister and nerve agents. The base chemicals and production procedures used are similar to those used in producing modern pesticides. The production process of blister agents is relatively simple, consisting of one or two steps, followed by production purification. Once purified, blister agents are highly stable and are able to be stored for many years. Nerve agents, on the other hand, are more complicated to make, requiring a series of steps.

Commercially available industrial chemicals are used as starting materials, or precursors, in the production of these chemical agents. Precursors are manufactured by a large number of countries, including a growing number of developing nations, making the diversion of chemicals to chemical agent production difficult to restrict. Most precursors have multiple high-volume legitimate uses.

The cost, availability of raw materials or precursors, stability, and difficulty of providing medical treatment to affected personnel are also important factors in any decision to manufacture chemical agents. Standard chemical agents such as the nerve agents tabun, sarin, VX, and soman and the blister agents mustard and Lewisite all have predictable properties and can be made by a number of well-known synthetic routes and thus have become the agents of choice for countries developing chemical warfare weapons. The technical challenge of preparing the agents varies. Mustard is the easiest to make and VX is the most difficult.

CW Weaponization

The weaponization of CW agents can be accomplished in numerous ways. The most common methods are the free-flight munitions that are fired at or dropped on a target. Weaponization can be in unitary or binary form, and the larger munitions can contain submunitions. Spray tanks can also be used to disseminate agents from an aircraft or from a ground-based aerosol generator.

Most conventional munitions can be modified to deliver lethal or nonlethal chemical agents. Typical chemical munitions may include aerial bombs, artillery rockets, artillery shells and mortar rounds, grenades, mines, missile warheads, and sprayers and spray tanks. These normally contain burster charges surrounded by the bulk-filled agent. The burster ruptures the munitions and causes the agent to be disseminated as a stream or cloud of small droplets.

Air- or ground-based aerosol generators can be used for more controlled dissemination of CW agents. Spray tanks can be used to disseminate agents from aircraft, and ground-based aerosol generators can be used to disseminate agents.

Chemical munitions usually fall into one of two categories: unitary or binary. A unitary munition contains the agent itself, whereas binary munitions contain two agent precursors that mix to form a warfare agent before or during flight. Unitaries are able to deliver more agent per munition, but binaries, because they contain the less-toxic precursors, are safer for storage.

CW agents can also be carried in submunitions or bomblets. The submunitions are ejected from the primary munitions some distance above the ground, landing on the ground in a random pattern and detonating, covering an area larger and more evenly than if deployed in bulk-filled munitions.

Terrorist organizations may choose dispersal methods that are less dependent on technology. Terrorists may utilize homemade explosives or spray devices. In the 1995 sarin attack on the Tokyo subway system, the Aum Shinrikyo cult used umbrella tips to puncture plastic bags containing the agent.

No matter what dispersal method is selected, however, the weather, terrain, and buildings will all have an effect on the chemical agent once it has been employed. Strong wind, heavy rain, or temperatures below freezing may reduce the effects of a chemical agent. After the attack, the weather will be of great importance for the respiratory risks expected at different distances from the targeted area. Similarly, weather conditions influence the effect of ground contamination.

After an attack, the primary chemical agent cloud will drift with the wind. Wind velocity will decide how long it will take the primary cloud to pass the targeted area. High wind velocity implies a short time of passage and thus fewer injuries to unprotected persons, whereas low wind velocity will lead to more injuries. Consequently, a weak wind may cause effects at greater distances than strong winds.

Wind velocity also naturally influences how fast the primary cloud moves. If the wind is gentle, then there are better opportunities for the victims to receive warning in time. In very weak winds, however, the gas cloud will not move very far. In addition, the wind direction varies widely in such situations, which is why a circular area must be alerted in an attack from CW agents.

The concentration in the primary cloud may also decrease in cold weather, particularly if the temperature is below freezing. This depends on a smaller amount of CW agents evaporating during dispersal, which also implies that the share of ground contamination will be greater.

Precipitation also reduces the concentration since some of the gas/aerosol is "washed" away by wet deposition. A major problem during the winter may be that contaminated snow on shoes and clothes is taken into shelters, vehicles, or buildings. Once in the warmth, the CW agent will evaporate and may result in concentrations of gas.

Light rain will cause ground contamination to be more dangerous since the pores in the soil become clogged and prevent the substance from penetrating down into the soil. Heavy rain, however, will flush off ground contamination, whereas heavy snow will cover it. In both cases, the contact risk is decreased.

Woodland and undulating terrains give shorter danger distances for the primary cloud since the wind will be exposed to greater turbulence. Woodland also absorbs a certain amount of gas and aerosol through dry deposition.

In or close to the target area, however, woodland, depressions, pits, and narrow streets may lengthen the effect of an attack. Gas and aerosol will be retained in these areas, particularly in situations of weak wind and stable stratification. The longest danger distances are obtained when the cloud passes over plains or lakes, or follows the contours of a valley.

The effect of a passing cloud of gas/aerosol will be delayed inside tents, buildings, and vehicles. Owing to the lower air exchange in such places, it will take longer for the cloud to penetrate. A certain amount of the CW agent will be taken up and bound on walls and other surfaces, which also contributes to decreases in concentration.

Consequently, it may be expected that there is a certain reduction in the effect of a passing cloud of gas. In ordinary buildings, the protection can be improved by closing doors and windows, turning off the ventilation, and sealing all cracks with tape.

Ground configuration is also important for the contact risk in ground contamination. A dry, hard, porous surface—that is, asphalt or concrete—will lead to fewer contact risks. On soft ground—for example, grass, moss, sand, or snow—it is easier to come into contact with CW agents that have penetrated the underlying surface. In dense woodland, the ground contamination is reduced and becomes uneven since the falling droplets are caught to some extent in the crowns of the trees. Terrain covered by bushes, on the other hand, may lead to major risks of contact.

Biological Warfare Agents

Biological warfare is defined as "the employment of pathogens or toxins to produce casualties in humans or animals and damage to plants or material." Pathogens are living organisms, whereas toxins are the poisonous by-products of living organisms.

Pathogens are disease-producing microorganisms such as bacteria, viruses, rickettsia, or fungi. They are either naturally occurring, or altered by mutation or genetic engineering.

Toxins are poisons produced by the metabolic activities of living organisms. Classic biological agents include botulinum toxin, smallpox, tularemia, Q fever, ricin, viral hemorrhagic fevers, and plague.

Biological agents are inherently more toxic than chemical agents on a weight-for-weight basis and provide broader coverage per pound of payload. Moreover, they are potentially more effective because most are naturally occurring organisms such as bacteria and viruses, which are self-replicating and have specific physiologically targeted effects, whereas chemical agents are manufactured chemicals that disrupt physiological pathways in a general way.

The potential range of materials used for biological agents is limitless, but a biological agent must possess a number of the following characteristics to be militarily useful.

Requirements:

1. Consistently produce a given effect such as death, disability, or plant damage
2. Be manufacturable on a large scale
3. Be stable under production and storage conditions, in munitions, and during transportation
4. Be capable of efficient dissemination
5. Be stable after dissemination

Desirable Characteristics:

1. Possible for the using forces to be protected against
2. Difficult for a potential enemy to detect or protect against
3. A short and predictable incubation period
4. A short and predictable persistency if the contaminated area is to be promptly occupied by friendly troops
5. Capable of:
   - Infecting more than one kind of target—for example, man and animals—through more than one portal of entry
   - Being disseminated by various means
   - Producing desired psychological effects

Biological weapons can be directly or indirectly employed against personnel, plants, animals, or material. Foods, especially uncooked food as in a salad bar; water supplies; or facilities can be rendered unsafe or unfit for use or consumption through contamination. People can be infected either directly by the employment of a biological agent or indirectly through secondary contamination from an exposed individual, as in smallpox and plague.

Both overt and covert attacks have the potential to cause large numbers of casualties. Most of the means of protecting against a chemical inhalant attack are largely effective against a biological inhalant attack, particularly the mask, although it is not a foolproof protection device. Recognize, however, that awareness that an attack is occurring is of the utmost importance, which is clearly problematic.

If the type of pathogen is not quickly determined and medical treatment is not readily available, very large numbers of casualties can be expected. From the time of an attack to the incubation period, zero casualties would report to medical personnel for treatment. Depending on the agent used, a peak in casualties would take place within a few days (for anthrax) or up to 2 months (for brucellosis) and could quickly overload medical personnel and facilities.

A biological attack may require treating those individuals who were exposed but have not yet exhibited symptoms. The extreme effects of a biological attack can be effectively mitigated. Bleach, hot soapy water, and sunlight are the best, most readily available means of decontamination.

Early detection is the key to mitigating the effects of a biological attack or incident. Rigorous implementation of the following activities is the "best practice" for preventing or detecting biological incidents:

- Water and food inspection programs
- Environmental monitoring programs
- Disease monitoring programs
- Proper physical protection facilities and personal protective equipment
- Forensic identification means
- Vaccine and diagnostic programs
- Training and awareness programs
- Good intelligence and event notification programs

Agent Production

Biological agents are a strategic as well as an operational threat. They can, depending on their intended use, cause lethal, disabling, contagious, or noncontagious, types of casualties.

Biological agents can be effectively employed against military targets such as headquarters, ship or aircraft crews, and troop concentrations, as well as civilian population centers. Missiles, aerosol generators, aerial line sprays, or covert means could be used to deliver the agents. Some agents, including microorganisms and toxins, are capable of widespread infection and debilitation.

To a country or terrorist group considering such a course of action, there are advantages to biological weapons. There are few reliable biological detection devices readily available, and biological agents are normally invisible to the human senses.

The delay in onset of symptoms could make it difficult to identify the time and place of attack. A biological attack could be attributed to a natural outbreak or epidemic, thus providing the attacking party with grounds for plausible denial.

There is little to distinguish a vaccine or pharmaceutical factory from a biological agent facility. Virtually all equipment, technology, and materials needed for biological agent production are dual use. The technical skills required to start and run a program are commensurate with the basic procedures of microbiology, with any required additional knowledge easily gained through training courses available from equipment suppliers or at scientific meetings.

Any nation with a modestly developed pharmaceutical industry has the necessary technical infrastructure to produce biological warfare agents. The equipment used to produce biological agents is dual use in nature, and a laboratory devoted to biological warfare research will generally be similar to a facility engaged in peaceful research activities.

The terrorist organization Aum Shinrikyo demonstrated the potential ease of weaponizing biological agents, in their case botulinum toxin, by cultivating strains of Clostridium botulinum from soil taken from one of their member's farms.

Recent advances in biotechnology research have increased the potential for developing new biological agents that have optimal weapon potential and reducing the time needed to develop new agents from basic research to mass production. In addition, the increasing technical complexity of biotechnological studies will make distinguishing between peaceful research and that directed toward biological warfare purposes even more difficult. This will allow an increasing number of nations to take advantage of advanced biotechnology developments.

It is not necessary for a country or terrorist organization to maintain large stockpiles of biological weapons because sufficient quantities of agents can be quickly produced using small starter cultures.

Biological Agent Weaponization

There are two methods of dissemination:

1. Line source: This technique is most effective using a dispersal means—a truck or air sprayer—moving perpendicular to the wind during an inversion, in which air temperature increases with altitude, holding surface air and pollutants down; incursions normally occur at dawn, dusk, or night.

TABLE 19-10 *Characteristics of selected pathogens*

Pathogen	Routes of Infection	Delivery	Untreated Mortality (%)	Incubation Period	Treatment
Anthrax	Skin, digestive, respiratory	Aerosol	Skin—less than 25%; respiratory—approaches 100%	1–4 days	Antibiotics (limited effectiveness after symptoms develop)
Plague	Vector, respiratory	Aerosol or vectors	Bubonic—50%; pneumonic—50%–90%	2–3 days	Antibiotics
Tularemia	Vector, skin, respiratory	Aerosol	30%–40%	1–10 days	Antibiotics
Q fever	Vector, respiratory	Covert or aerosol	Less than 1%	14–26 days	Antibiotics
Brucellosis	Digestive, respiratory	Aerosol	Less than 6%	5–21 days	Antibiotics
Smallpox	Direct contact	Covert	30%–40%	7–14 days	Supportive care only
Viral hemorrhagic fevers (Ebola, Marburg)	Direct contact	Aerosol	40%–90%	4–21 days	Supportive care only

TABLE 19–11 *Characteristics of selected toxins*

Toxin	Natural Source	Rate of Action	Dose of 50% Mortality Rate*	Effect
Botulinum	Clostridium botulinum bacteria	1–12 hours	0.0003 to 0.01 μg/kg	Dilated pupils, double vision, dry mouth, paralysis
Tetanus	Clostridium tetani bacteria	1–12 hours	0.0025 μg/kg	Muscle spasms, frequently in the jaw muscles
Palytoxin	Palythoa soft corals	5 minutes	0.08	Muscle contractions, heart irregularities, rigid paralysis
Batrachotoxin	South American frog	5 minutes to 1 hour	0.1 to 2 μg/kg	Loss of coordination, numbness, headache, irregular heart rate, respiratory paralysis
Ricin	Castor bean	5 minutes to 1 hour	3.0 (oral) μg/kg	Nausea, vomiting, cramps, bloody nose, diarrhea, difficulty breathing, twitching
Saxitoxin	Shellfish	5 minutes to 1 hour	5–12 (oral) μg/kg; 1 (aerosol) μg/kg	Tingling, numbness, weakness, flaccid (limp) paralysis
Tetrodotoxin	Puffer fish	5 minutes to 1 hour	30 (oral) μg/kg	Vomiting, tingling, numbness, lack of muscle control, loss of voice, paralysis of arms and legs
Tricothecene (T2) mycotoxin	Fusarium mold on infected grain	1 to 12 hours	50–240 (aerosol) μg/kg	Itching, tingling, vomiting, hemorrhaging, bloody diarrhea
Staphylococcus enterotoxin Type B (SEB)	Staphylococcus aureus bacteria	1 to 12 hours	200 (aerosol) μg/kg	Vomiting, cramps, nausea, diarrhea, severe weakness

* Lethal doses based on μg per kilogram of body weight

2. Point source: This technique uses small bomblets deployed in a saturation mode. The saturation technique overcomes the meteorological requirements for line source dissemination. Agents may be introduced into buildings' heating and ventilation air-conditioning systems or via food or water contamination. Small packages or envelopes may also be used to disperse the agent.

Many of the technologies and processes used for weaponizing chemical warfare agents can also be applied to biological warfare agents. Aerial bombs, rockets, artillery, missile warheads (with or without submunitions), and spray systems can be used to deliver biological warfare agents. The method of delivery depends on operational objectives, capabilities, and strategic and tactical doctrine.

Both aerosol attacks and food and water contamination attacks have the potential to cause large numbers of casualties. Although decontamination is individually useful, area decontamination after a biological attack may not provide further reduction in casualties, depending on the agent involved.

A special problem is associated with contamination that tends to settle in basements and other low areas, rubble piles, and similar collections of debris or into porous surfaces. This concentration could extend the lethality period of biological agents.

Tables 19-10 and 19-11 represent the lethality of a representative sample of biological agents.

[i] Glasstone

[ii] Glasstone

[iii] Glasstone

[iv] Glasstone

[v] Glasstone

[vi] Glasstone

[vii] Glasstone

[viii] Glasstone

Sources

Chapter 1

"Fundamentals of Naval Weapons Systems." *Federation of American Scientist.* 30 September 2002, http://fas.org/man/dod-101/navy/docs/fun/index.html.

Chapter 2

Electronics Technician Volume 7—Antennas and Wave Propagation. NAVEDTRA 14092. Navy Education and Training Professional Development and Technology Center, 1995.

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Chapter 3

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

Gurvine, Jeff R., and Edwin G. Strauss. *Fundamentals of Tactical Missiles.* Tucson, Ariz.: Raytheon Missile Systems Company, 1997.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Stimson, George W. *Introduction to Airborne Radar.* 2nd ed. Raleigh, N.C.: SciTech, 1998.

Chapter 4

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Sullivan, Roger J. *Radar Foundations for Imaging and Advanced Concepts.* Raleigh, N.C.: SciTech Publishing, 2004.

Chapter 5

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Chapter 6

Budd, Nicholas. "Electronic Warfare." Lecture notes, United States Naval Academy, October 2002.

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

"Fundamentals of Naval Weapons Systems." *Federation of American Scientist,* 30 September 2002. http://fas.org/man/dod-101/navy/docs/fun/index.html.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Chapter 7

Boreman, G. D., and E. L. Dereniak. *Infrared Detectors and Systems.* New York: John Wiley and Sons, 1996.

Butler et al. *EOS Science Steering Committee Report.* Vol. 2, *From Pattern to Process: The Strategy of the Earth Observing System.* Washington, D.C.: NASA, 1987.

Driggers, Ronald G., and P. E. Cox. *Introduction to Infrared and Electro-Optical Systems.* Boston: Artech House, 1999.

"Early Attempts at Night Vision Technology." *U.S. Army Night Vision and Electronic Sensors Directorate.* http://www.nvl.army.mil/about/index.shtml.

Encarta Encyclopedia Deluxe. 1999 ed. s.v. "Galileo." http://encarta.msn.com.

Gurvine, Jeff R., and E. G. Stauss. *Fundamentals of Tactical Missiles.* Tucson, Ariz.: Raytheon Missile Systems Company, 1998.

Hudson, Richard D., Jr. *Infrared Systems Engineering.* New York: John Wiley and Sons, 1969.

Palmer, James M. "Radiometry and photometry FAQ." *Optical Sciences Center University of Arizona.* http://www.optics.arizona.edu/Palmer/rpfaq/rpfaq.htm.

Senich, Peter R. *U.S. Marine Corps Scout Sniper: World War II and Korea.* Boulder, Colo.: Paladian Press, 1993.

Wolfe, William, and George Zissis, eds. *The Infrared Handbook.* Washington, D.C.: Office of Naval Research-Department of the Navy, 1985.

Chapter 8

Anderberg, Bengt, and Myron L. Wolbarsht. *Laser Weapons: The Dawn of a New Military Age.* New York: Plenum Press, 1992.

Driggers, Ronald G., and P. E. Cox. *Introduction to Infrared and Electro-Optical Systems. Boston*: Artech House, 1999.

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

Gurvine, Jeff R., and E. G. Stauss. *Fundamentals of Tactical Missiles.* Tucson, Ariz.: Raytheon Missile Systems Company, 1998.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Hong, J. S., D. A. Levin, R. J. Collins, J. Emery, and A. Tietjen. "Comparison of Atlas Ground Based Plume Imagery with Chemically Reacting Flow Solutions." AIAA Paper No. 98-2537, 32nd Thermophysics Conference, June 1997.

Hudson, Richard D., Jr. *Infrared Systems Engineering.* New York: John Wiley and Sons, 1969.

Pratt and Whitney Aircraft, Division of United Aircraft Corporation. East Hartford, Conn.

USN/USMC. *Helicopter Night Vision Device (NVD) Manual.* 6th ed., 2002.

Chapter 9

Coppens, Alan B., James Sanders, and Harvey Dahl. *Introduction to the Sonar Equations.* Monterey, Calif.: Naval Postgraduate School, 1982.

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

"Fundamentals of Naval Weapons Systems." *Federation of American Scientist,* 30 September 2002. http://fas.org/man/dod-101/navy/docs/fun/index.html.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Kinsler, Lawrence E., et al. *Fundamentals of Acoustics.* New York: Wiley, 1982.

"Shallow Water Acoustics." Naval Pacific Meteorology and Oceanography Center. 20 February 2004. http://metoc.npmoc.navy.mil/KBay/shallowacoustics.com.

"Ship's Silencing Program." *Federation of American Scientist.* 30 September 2002. http://www.fas.org/man/dod-101/navy/docs/swos/stu2/NEWIS9_7.html.

Urick, Robert J. *Principles of Underwater Sound.* New York: McGraw-Hill, 1983.

U.S. Navy. FASOTRAGRULANT. *Air ASW Oceanography.* D-5B-0010. Norfolk, Va.: Naval Air Station, 1988.

———. Naval Oceanographic Office. *Fleet Oceanographic and Acoustic Reference Manual.* MS39522-5001. 1992.

———. Naval Ordnance Systems Command. *Weapons Systems Fundamentals.* NAVORD OP 3000, vol. 1, 1st rev. Washington, D.C.: GPO, 1971.

Chapter 10

Coppens, Alan B., James Sanders, and Harvey Dahl. *Introduction to the Sonar Equations.* Monterey, Calif.: Nava/Postgraduate School, 1982.

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

"Fundamentals of Naval Weapons Systems." *Federation of American Scientist,* 30 September 2002. http://fas.org/man/dod-101/navy/docs/fun/index.html.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Urick, Robert J. *Principles of Underwater Sound.* New York: McGraw-Hill, 1983.

U.S. Navy. Naval Oceanographic Office. *Fleet Oceanographic and Acoustic Reference Manual.* MS39522-5001. 1992.

———. Naval Ordnance Systems Command. *Weapons Systems Fundamentals.* NAVORD OP 3000, vol. 1, 1st rev. Washington, D.C.: GPO, 1971.

Chapter 11

Albers, Vernon M. *Underwater Acoustics Handbook.* Lsnvsdyrt: Pennsylvania State University, 1960.

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

"Fundamentals of Naval Weapons Systems." *Federation of American Scientist,* 30 September 2002. http://fas.org/man/dod-101/navy/docs/fun/index.html.

Grace, Donn O. *Sonar and Anti-Submarine Warfare.* Escondido, Calif.: Sage, 1999.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

Horton, Warren J. *Fundamentals of Sonar.* Annapolis, Md.: United States Naval Institute, 1957.

Kinsler, Lawrence E., et al. *Fundamentals of Acoustics.* New York: Wiley, 1982.

Tucker, D. G., and B. K. Grazey. *Applied Underwater Acoustics.* London: Pergamon, 1966.

Urick, Robert J. *Principles of Underwater Sound.* New York: McGraw-Hill, 1983.

U.S. Navy. Naval Material Command. *Principles and Applications of Underwater Sound.* NAVMAT P-9674. Washington, D.C.: GPO, 1968.

———. Naval Ordnance Systems Command. *Weapons Systems Fundamentals.* NAVORD OP 3000, vol. 1, 1st rev. Washington, D.C.: GPO, 1971.

Chapter 12

F414-GE-400, GE Aircraft Engines. Customer Technical Education Center.

"Fundamentals of Naval Weapons Systems." *Federation of American Scientist.* 27 May 2004. http://fas.org/man/dod-101/sys/ship/weaps/mk-45.htm.

Nise, Norman. *Control Systems Engineering.* Redwood City, Calif.: Benjamin Cummings Publishing, 1995.

Sloan, Bliss. "What Is a Flyball Governor?" 27 May 2004. www.eagleridgetech.com/sloan_on_flyball_governor.html.

Smith, Billy R. *Weapons Systems Engineering.* Dubuque, Iowa: Kendall Hunt, 1993.

Chapter 13

Corse, Carl D., Jr. *Introduction to Shipboard Weapons.* Annapolis, Md.: United States Naval Institute, 1975.

Frieden, David R., ed. *Principles of Naval Weapons Systems.* Annapolis, Md.: United States Naval Institute, 1985.

"Fundamentals of Naval Weapons Systems Ring Laser Gyro." *Federation of American Scientist.* 30 September 2002. http://fas.org/man/dod-101/navy/docs/fun/rlg.htm.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems.* Dubuque, Iowa: Kendall Hunt, 2000.

U.S. Navy. Naval Education and Training Support Command. *Principles of Naval Ordnance and Gunnery*. NAVEDTRA 10783-C. Washington, D.C.: GPO, 1974.

———. Naval Ordnance Systems Command. *Weapons Systems Fundamentals*. NAVORD OP 3000, vol. 1, 1st rev. Washington, D.C.: GPO, 1971.

Chapter 14

Frieden, David R., ed. *Principles of Naval Weapons Systems*. Annapolis, Md.: United States Naval Institute, 1985.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems*. Dubuque, Iowa: Kendall Hunt, 2000.

U.S. Navy. Naval Ordnance Systems Command. *Weapons Systems Fundamentals*. NAVORD OP 3000, vol. 1, 1st rev. Washington, D.C.: GPO, 1971.

Chapter 15

Federation of American Scientist, 30 January 2002. http://fas.org/man/dod-101/sys/dumb/index.

Frieden, David R., ed. *Principles of Naval Weapons Systems*. Annapolis, Md.: United States Naval Institute, 1985.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems*. Dubuque, Iowa: Kendall Hunt, 2000.

U.S. Navy. Naval Ordnance Systems Command. *Weapons Systems Fundamentals*. NAVORD OP 3000, vol.1, 1st rev. Washington, D.C.: GPO, 1971.

Chapter 16

Frieden, David R., ed. *Principles of Naval Weapons Systems*. Annapolis, Md.: United States Naval Institute, 1985.

"Fundamentals of Naval Weapons Systems." *Federation of American Scientist*, 30 September 2002. http://fas.org/man/dod-101/navy/docs/fun/index.html.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems*. Dubuque, Iowa: Kendall Hunt, 2000.

Harris, Tom. "How a Sidewinder Missile Works." *Howstuffworks*, December 2002. www.howstuffworks.com/sidewinder4.htm.

U.S. Navy. Naval Ordnance Systems Command. *Weapons Systems Fundamentals*. NAVORD OP 3000, vol. 1, 1st rev. Washington, D.C.: GPO, 1971.

Chapter 17

Akhavan, Jacqueline. *The Chemistry of Explosives*. Letchworth, England: Royal Society of Chemistry, Turpin Distribution Services Limited, 1998.

Cooper, P. W., and S. R. Kurowski. *Introduction to the Technology of Explosives*. New York: VCH Publishers, Inc., 1996.

Daratony, A. J. *Blasting and Explosives Safety Training Manual: A Guide for the Industrial Blaster*. Long Beach, Calif.: Eastside Service Company, 1985.

Frieden, David R., ed. *Principles of Naval Weapons Systems*. Annapolis, Md.: Naval Institute, 1985.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems*. Dubuque, Iowa: Kendall Hunt, 2000.

Sickler, Robert E. *Explosive Principles: An Essential Guide to Understanding Explosives and Detonations*. Boulder, Colo.: Paladin Press, 1992.

Small Arms and Special Warfare Ammunition Technical Manual. NAVSEA SW010-AD-G7P-010, 2nd rev., Change B. Washington, D.C.: Naval Sea Systems Command, 30 September 1999.

Chapter 18

Federation of American Scientist, 18 January 2002. http://www.fas.org/man/dod-101/sys/land/docs/ordtech. pdf.

Federation of American Scientist, 18 January 2002. http://www.fas.org/man/dod-101/sys/land/docs/warheads.pdf.

Federation of American Scientist, 30 January 2002. http://fas.org/man/dod-101/sys/dumb/index.

Federation of American Scientist, 30 January 2002. http://fas.org/man/dod-101/sys/smart/index.html.

Frieden, David R., ed. *Principles of Naval Weapons Systems*. Annapolis, Md.: United States Naval Institute, 1985.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems*. Dubuque, Iowa: Kendall Hunt, 2000.

Keevin, T. M., and G. L. Hempen. *The Environmental Effects of Underwater Explosions with Methods to Mitigate Impacts*. U.S. Army Corps of Engineers, St. Louis District, August 1997.

U.S. Navy. Naval Ordnance Systems Command. *Weapons Systems Fundamentals*. NAVORD OP 3000, vol. 1, 1st rev. Washington, D.C.: GPO, 1971.

Walters, W. P., and Zukas, J. A. *Fundamentals of Shaped Charges*. New York: Wiley, 1989.

Chapter 19

Deputy Assistant to the Secretary of Defense for Chemical and Biological Defense. U.S. Department of Defense. *Chemical and Biological Defense Primer*. Washington, D.C., 2001.

Glasstone, Samuel, and P. J. Dolan. U.S. Department of Defense. *The Effects of Nuclear Weapons* 3rd ed. Washington, D.C.: GPO, 1977.

Hall, Joseph B., ed. *Principles of Naval Weapons Systems*. Dubuque, Iowa: Kendall Hunt, 2000.

Index

Page numbers in bold include illustrations and tables

absorption, 15, 119–21
absorption bands, 120
absorption coefficient, 183
absorption lines, 120
absorption losses, 183
absorption vs. frequency, **184**
absorptivity, 111
acceleration-integration S&A device principle, **315**
acceleration sensing, 315
accelerometers, 299–**300**
accuracy of homing, 307
acoustic countermeasures, 210–12
acoustic intensity, 181
acoustic intercept receivers, 210
acoustic jammers, 212
acoustic mine mechanism, **322**
acoustic mines, 211
acoustic reference conversion factors, **156**
acoustic towed decoys, 211
acoustic variability, **170**
acoustic waves, 154
acquisition gate, 77
active acoustic torpedoes, 322
active beamforming, **197**
active deception repeaters, 93–94
active electronically scanned array (AESA), **62–63**
active equation system, **188**
active homing, 293, 306
active IR searchlight, 100
active/passive decoys, 93
active sensors, 153
active sonar equation, 187–88
active sonar system, **196**
active source level, 180
active systems, 196–200
  beamforming processor, 197–98
  display, 198
  Doppler, 198–200
  duplexers, 198
  receivers, 198
  synchronizers, 198
  transducer array, 196–97
  transmitters, 197
active tracking radar, 81–82
active tracking technique, 82
actuation distance, 323
adaptive beamforming, 197
additional effects of nuclear explosions, 368–69
addition and subtraction, 228
ADM-141 TALD (tactical air= launched decoy), 93
ADT processor, 76, **77**
advanced gun system long-range land attack projectile, **292**
Aegis system radar, 61
Aegis weapons systems, 275
aerodynamic control surfaces (fins), 288
aerodynamic drag, 289–90
aerodynamic forces, 282, **290**
aerodynamic heating, **130**
aerodynamic heating signatures, 130–32
  armored vehicles, 131–32
  cruise missile targets, 130–31
  missile signatures, 131
  sea-based sources, 132
aerodynamic stabilization, 290
aerosol scattering, 121, 122
AESA transmit/receive module, **62**
afterwinds, 367
agent delivery, 383–84
agent production (biological), 389–90
agent production (chemical), 387
AGM-88 high speed antiradiation missile (HARM), 94
AH-1W night targeting system (NTS) FLIR, 139
AIM-9X, 138
airborne early warning (AEW) radar, 40
airborne launcher rails, 269
air burst, 369–70
aircraft
  ejection launchers on, 268
  exhaust plume targets and backgrounds, 126–30
  gun launchers for, 266
  signatures, 126–27
air glow, 116
air laid moored mine, **296**
air launched cruise missiles (ALCMs), 2, 281
air target trajectories vs. surface targets, **244**
alarm signals, 36
all-aspect weapons, 125
alpha particles, 372
amatols, 338
ambient fuzes, 317
ambient noise, 175
ambient noise-limited situation, 187–88
ambiguity resolution of crossing tracks, **79**
ambiguous range and high pulse repetition rates, 50
ammonium nitrate, 338
amplitude, 7
amplitude changes of received energy with target position, **74**
amplitude comparison monopulse, 307
amplitude modulation (AM), 44
ANFO, 338
angle deception, 92
angle-of-arrival measurement capability, 84
angle of attack, 284, 285, 287
angle of departure, 246
angle of fall, 246
angle of incidence, 10
angle of inclination, 163
angle of reflection, 10
angle-sensitive circuitry, 92
angles of rotation, 256
angles of variation, 208
angle tracking, 72
angle tracking circuit deception, 92
angular magnification, **135**
angular resolution, 51–52
annular blast fragmentation (ABF) warheads, 352
anomalous detection, 41
AN/PVS-5, **149**
AN/SAY-1 camera systems, 132
AN/SLQ-25 NIXIE, 211
AN/SPS-33 radar, 62
AN/SPS-48 series radars, 58
AN/SPY-1 radar, 62
AN/SYS-2 IADT systems, 82
antenna beamforming, 18–19
antenna beam theory, 12
antenna gain, 34–35
antenna gain and directivity, 35
antennas
  broadside arrays, 21
  design of, 96
  electromagnetic waves, 16–17
  endfire arrays, 21

antennas (*Continued*)
functions of, 29
generation of electromagnetic waves, 15–16
linear antenna arrays, 20–21
Maxwell's theory, 15
parasitic elements, 21
quasi-optical reflectors, 19–20
wave polarization, 17
antenna-to-antenna coupling, 41
anti-ES countermeasures, 97
antijamming (AJ) fixes, 96
antilogs, 158
antipersonnel warheads, 357
antiradiation (ARM) weapons, 89, 94, 307
AN/TPS-59 three-dimensional radar, 61
AN/WLR-2, 210
aperture, 12, 107
arcing, 37
arithmetic selectivity, 85
armament datum line, 260
armored vehicles, 131–32, 352
armor piercing (AP), 290
armor-piercing, fin-stabilized, discarding-sabot (APFSDS) rounds, 355
armor-piercing fuzes, 325
armor-piercing projectile, 290
armor-piercing warheads, 355–56
array with elements radiating in phase, 54
arsenicals, 387
artificial illumination, 116
A-scan display, 29, 198
A-scope presentation, 48
Atlas IIA rocket, 131
Atlas rocket, 131
atmosphere
parameters as a function of height, 119
vertical temperature profile of, 119
atmospheric gases transmission, 120
atmospheric particles, 122
atmospheric refraction of space waves and radar horizon, 24
atmospheric structure, 118–19
atmospheric trajectories vs. vacuum, 245–46
atmospheric transmission, 121, 123
atmospheric transmission windows, 120
attenuation, 15, 345
attitude control systems, 298
automatic detection and tracking (ADT), 76
automatic gain control (AGC), 90
automatic-tracking system, 68
autopilot, 294
availability and cost, 332
Avogadro's law, 337
axial-flow turbojet, 280
azimuth, 29
azimuth angle, 59, 60
azimuth tracking, 69–73
conical scan, 69–71
COSRO (conical scan on receive only), 71
inverse conical scan, 71
lobe on receiver only (LORO), 71
lobe-switching, 69
monopulse, 71–73
sequential lobing, 69
azimuth vs. cross section of tactical bomber-sized aircraft, 38

background limited systems, 145
background noise, 35
backscatter from different shapes, 39
baffles, 197
ballistic procedure
considering gravity only, 244
input/output relationship of, 242
ballistics procedure, 247
bandpass, 84
bandpass filter
defined, 84
pulse-Doppler (PD) radar as, 46
bandwidth, 36–37, 50
bandwidth factor, 144
barrage jamming, 89
Basic Point Defense Missile System, 66
bathythermograph (BT), 160
beam axis, 58
beam axis and target signal strength, 33
beam depression from phase lag, 58
beam direction changes and offset feed, 72
beamforming, 19
beamforming processor, 197–98, 201–2
beam patterns directivity, 194
beam positioning in relationship to the boresight axis, 56
beam power distribution pattern, 185
beam radiation pattern, 19
beam-rider guidance system, 305
beam-rider method, 305
beam riding, 303–4
beam splitting, 76
beam steering control (BSC), 62
beam steering methods, 55–56
beamwidth
flat plate halfpower point, 194
of fragmenting payload, 351
measurements of, 33
as measure of angular resolution of radar, 19
quasi-optical reflectors for narrowing, 19
tactical choices regarding, 32–34
bearing, 29
bearing ambiguity, 205
bearing and range display, 29
bearing time history (BTH) display, 202
beat frequency, 44
Beaufort Wind Scale and sea states, 176
becquerel, 372
bed plate, 174
beehive round, 357
Beer-Lambert law, 123
Beer's law, 123
beneath keel explosion, 359
Bernoulli hump, 207
best depth, 166
beta particles, 372
binary chemical munitions, 388
binary warheads, 356
binding energy, 361–62
bins, 342–43
biological agent production, 389
biological agents vs. chemical agents, 383
biological agent weaponization, 390–91
biological luminescence, 115
biological noise, 176–77
biological warfare agents, 389–91
agent production, 389–90
biological agent weaponization, 390–91
biological warheads, 356
Biological Weapons Convention, 382
bipropellants, 279
bistatic reflection, 306
bistatic sonar, 207
bivariate distribution, 343
blackbodies
and quantum theory, 110
spectral exitance of, 112
spectral radiant exitance of, 111
and temperatures, 113
blackbody radiation, 110
black boxes, 86
blackout, 368
black powder, 240
blade-tip cavitation, 172
blast effects damage criteria, 350
blast warheads, 346–50
blast wave, 366–67
blast wave effects at a given distance from blast center, 347
blast wind, 346, 367
blind speeds, 49, 58, 96
blinking, 92
blip enhancers, 93–94
blister agents, 385, 386–87
block diagram addition, 228
block diagrams
closed-loop system, 229
with controller, 233
error time response, 233
forward path-amplification and unity feedback, 230
with negative feedback, 227
negative feedback, 231
negative feedback loop, 230
open-loop speed control, 227
open-loop with forward-path amplification, 229
position feedback, 231
positive feedback loop, 229
with proportional control, 233
rate and position feedback, 231
spring, 228
unity feedback, 230
vehicle speed control system, 232
block simplification, 229
blood agents, 386
blooming, 148
boat tail, 289
body armor, 352
body-fixed RF interferometer, 307
Boeing 707, 129
bolometer configuration, 142
bolometers, 141–42
boost (initial) phase, 300–301
boost-glide weapons, 289
boresight axis, 56

boresighting equipment, **260**
bottom absorption and scattering, 184
bottom bounce, 168
bottom bounce propagation, **168**
bottom bounce reception, **205**
bottom mines, 295
bottom reverberation, 184
bourrelet, 289
brisance, 334
broadband display, 202
broadband source, **170**
broadband sources, 170, 180
broadside arrays, **21**
bulk modulus, 159
Bunsen burner spectral radiance, **128**
buoyancy, 292
burning rate, 240, 278
burnthrough, 88
burst types, 369–70

canard control, 287–88
cancellation circuit, **48**
cancellation of echoes, **48**
canister launchers, 270–72
cannon launchers, 262
Captor mine, 295
carbon dioxide, 126–27
carrier frequency, 27, 32
Cartesian coordinates, **238**
cartridge, 291
case guns, 291
catapult ramming, **274**
cavitation, 172–**73**, 196
cavity effect, 353
celestial guidance, **309**
celestial reference, 309
center of gravity (C.G.), 287
center of pressure (C.P.), 282, 287
centigray, 373
centisievert, 373
central frequency bandwidth vs. pulse width, **37**
ceramic transducers, 193
CFAR receivers, 36
CH-60 Blackhawk IR signature, **127**
chaff, 93, 96
chaff warheads, 357
chain ramming, **274**
chain reaction, 363
characteristic equation, 219
characteristic impedance, 156
charge-coupled devices, 142–43
charge-to-metal ratio, 350, 351
Charles's law, 337
chemical, biological, and radiological (CBR) weapons, 361
chemical agents
  vs. biological agents, **383**
  defined, 385
chemical and biological weapons (CBW), 382–84
  CBW characteristics, 382–84
  delivery methods of, 383–84
  symptoms of, 384
chemical composition, 333
chemical explosives
  see explosives
chemical explosives vs. fuel-air-explosive (FAE), 356
chemical-type guns, 269
chemical warfare agents, **385**
  agent production, 387
  blister agents, 386–87
  blood agents, 386
  choking agents, 386
  CW weaponization, 387–88
  nerve agents, 385–86
chemical warfare weaponization, 387–88
chemical warheads, 356
Chemical Weapons Convention, 382
chirp pulse, 95
choking agents, 386
circular error possible (CEP), 341–45
  discrete and continuous distributions, 342–44
  random error, 344
  and random error circles, **345**
  systematic error, 344–45
circular polarization, 18
circular polarized antenna, **18**
classification, 1, 3
clock mechanism, 324
closed-loop systems, 229–30
closed-loop transfer function (CLTF), 230–31
closing velocity, 320
cluster bomb separation, **357**
cluster bomb units (CBUs), 318, 357
clutter, 45, 76, 123
clutter cell, 77
clutter map, 77
coherency, 6
coherent energy, 6
coherent light, 133
coherent waves, 32
coil gun, 268
collimation of multiple-beam fire control radar, **259**
collision path, 303
combat information center (CIC), 1
combination fuzing, 323
combination of multiple frequencies, **14**
combination radar systems, 45–51
  MTI mode, 48–49
  pulse-Doppler radar mode, 46–48
  velocity discrimination, 46
  velocity limitations of, 49–50
combustion chamber, 279
command fuzes, 325–26
command guidance method, 304
command guidance systems, **305**
communication bands, **22**
compensator, 233
composite explosives, 338, **338**
composite guidance systems, 308
composite multipath representation, **169**
composite system controls, 235–36
compressional wave train, 154
Compton scattering, 371
computation of required phase relationships, 59–62
conical scanning, **69**–71
conical scan systems, signal-to-noise ratio (SNR) of, 73
conical scan tracking, 305
conservation of energy, 111
constant-bearing path, **303**
constant-false-alarm rate (CFAR), 36, 96
constant-false-alarm rate (CFAR) receivers, 96–97
constant preset, 301–2
constant preset guided path, **302**
constructive interference, 13–14, 54, 55
contact fuzes, 288
contact sensing devices, 316
continuous distribution functions, 343
continuous-rod bundle sections, **353**
continuous-rod warheads, 352–**53**
continuous wave (CW)
  antenna gain, 34–35
  bandwidth, 36–37
  beamwidth, 32–34
  carrier frequency, 32
  pulse repetition frequency, 31–32
  pulse shape and width, 30–31
  radar cross section (RCS) of targets, 37–39
  receivers, 35–36
  scan rate, 34
  transmitter power, 37
continuous wave (CW) radar
  block diagram, **42**
  described, 42–44
  frequency modulated CW radars, 44–45
  history and early use of, 26
  limitations of, 43–44
  modulation of, 44
continuous-wave Doppler deception, 92
control, 274–75
control and system guidance, 293–94
control guidance, 304–6
controller performance summary, 235
controller types, 233
control surface placement, **287**
control system diagram for full authority digital electronic control (FADEC), **214**
control systems
  defined, 213
  error, 231–33
  history of, 213–14
  system controllers or compensators, 233–36
control systems block diagrams
  addition and subtraction, 228
  block simplification, 229
control systems modeling
  electrical systems, 216–17
  Laplace transforms, 217
  mathematical representation, 214–15
  mechanical systems, 215
  rotational systems, 216
  system response, 219–20
  translational systems, 215–16
conventional recoil vs. soft recoil, **266**
convergence zone, 167–68
convergence zone propagation, **167**
conversion factors, acoustic reference, **156**
coordinate conversion and transformation, **255**

coordinate system conversion, 254
coordinate systems
  coordinate system conversion, 254
  coordinate transformation (reference frame rotation), 255–57
  cylindrical coordinate systems, 254
  rectangular coordinate systems, 254
  and reference frames, relationship between, **256**
  reference frame translation, 257–58
  spherical coordinate systems, 253–54
coordinate transformation (reference frame rotation), 255–57
Coriolis effect, 246–47
corona, 37
corrosion, 242
cosine wave, **9**
cosmic radiation, 375
COSRO (conical scan on receive only), 71
countermeasures, 210–12
counterrecoil, 264
critical angle, 165
critical frequency, 23
critically damped response, 220
critically damped time response, **221**
critical mass, 363
critical ray, 165
crosseye (phase front distortion) and TWS radar, 92
crossing tracks ambiguity resolution, **79**
cross-layer detection, 168
cross-range resolution, 52
cross section
  vs. azimuth, **38**
  of gun propellant configurations, **291**
  of spherical wave spreading, **10**
crosswind, 246
cruciform grain, 278
cruise missile, **130**
cruise missile targets, 130–31
crustaceans, 176
crystal stack for transducer, **193**
crystal transducers, 192
curie, 372
cyanogen chloride, 386
cyclosarin, 385
cylinder expansion test, 334
cylinder fragmentation test, 334
cylindrical coordinate systems, 254
cylindrical spreading, **182**
cylindrical transducer array (declination angle), **197**

damage
  categories of, 342
  probability of, 341
damage criteria
  for blast effects, **350**
  for fragmentation effects, **352**
damage description, 341–42
damage potential, 321
damage quantification, 341–42
damage volume, 345
dampers, 299
damping ratio, 220
deception jamming, 90–94
decibels, 157–59
decibels, nomogram to combine, **158**
decomposition products, **339**
decomposition reactions, **336**
deep isothermal layer, 162
deep-sea speed profile divided into layers, **162**
deep sound channel, 162, **167**
degaussing, 208
degressive burning grains, **240**
De Laval nozzle, **279**
delay, 332
delay fuzes, 325
delay line canceller, 48
delay mechanisms, 316
delivery methods, 383
denial jamming, 87–90
density of loading, 334
deperming, 208
depth excess, 167–68
depth of focus, **136–37**
depth of penetration, 354
destruction or neutralization, 2
destructive interference, 13–14, 54, 94
detection and identification, 1
detection of ships, 64
detection ranges, **151**, 179
detection threshold (DT), 179–80, 198
detector responsibility, 143
detectors, 138–43
detectors and scanners, **138**
detect to engage sequence, 1, **2**, 3
determination of critical diameter, 334
deterministic effects, 375
detonation pressure (Chapman-Jouget) test, 334
detonation sequence and explosive train, 331–32
detonator, 331
deuterium with tritium, 362
DICASS (directional command actuated sonobuoy system), 206
DIFAR (directional finding acoustic receiver), 206
differential GPS (DGPS), 312
diffraction
  described, 11–12
  of ocean swell around a breakwater, **11**
  of a radar wavefront around an obstruction, **12**
  through a slit, **12**
diffraction loading, **348**, 359, 376
diffuser, 279, 280
diffuse reflectors, 106
diffusion from an irregular surface, **10**
digital scene matching area correlator (DSMAC), 310
diphosgene, 386
dip-needle influence mines, 321
dipole antennas
  electromagnetic energy field from, 18–19
  fields in space around, 17
  generation of an electric field about, **15**
  generation of an magnetic field about, **16**
dipping sonar, 203
direct detection receivers, **84–85**
directionality, **195**
directional noise, 185
directional reflectivity, **106**
direction determination, 195
directive gain vs. power gain, 35
directivity
  and antenna gain, 35
  of beam patterns, 194
  and carrier frequency, 32
  defined, 34
  quasi-optical reflectors and, 19
directivity factor, 185
directivity index (DI), 185–87, **186**, 204, 205
direct line-of-sight, 24–25
direct path, **205**
direct-path propagation, **166**
discrete and continuous distributions, 342–44
discrete distributions, 343
discrete Fourier transforms (DFT) method, 47
display, 29, 84, 198
distance to which effects will be felt, 378
distributed isolation material (DIM), **175**
distribution functions, 343
distribution of five rounds, **343**
distribution of fragments, **351**
distributions, behavior of, 343
disturbance, 227
Doppler, 198–200
Doppler drift, 32
Doppler effect, 252
Doppler filter bank, 47
Doppler filter bank bypass output, 47
Doppler filter bypass, 47
Doppler frequency gates, **46**
Doppler frequency resolution, 53
Doppler principle, 319
Doppler radars, 50, 96
Doppler shift, 42, 43, 53
dose on target, 384
doses inhaled or absorbed, 384
double-integrating accelerometers, 299
down Doppler, 200
downslope sound propagation, 169
downwash, 285
downwind dispersal, 384
drag, 244–46
drag characteristics, **285**
drag coefficient, 285
drag coefficient vs. mach number, **290**
drag equation, 284, 346
drag forces, 244, **245**, 290
drag loading, **348**, 359, 376
D region, 23
drift, 246
drifting mines, 295
dry land drowning, 386
dual-beam angular output determination, **72**
dual beams, **72**
dual-use items, 382
duplexers, 29, 198
duty cycle (DC), 27, 42
dynamic overpressure as shock wave, 367
dynamic pressure
  from blast winds, 367
  and drag coefficient, 285
  drag loading and, 348–49
  as function of maximum wind speed, **367**
  and lift, 284

vs. time relationship of blast wave, **346**
for TNT, **349**
dynamic volumetric gate positioning, **81**

EA antennas, **91**
EA ladder, **98**
early range gates, 75
earth rate, 247
EA techniques, **93**
echo enhancers, 93–94
echoes, cancellation of, **48**
effective angle of attack, 285
effective antenna aperture, 35, 39
effective aperture, 52
effective earth radius, **25**
effective pressure, 155
ejector launchers, 262, 267–68
elasticity, 159
electric actuators (motors), 288
electrical boresighting, 260
electrical circuit, **216**
electrically damped accelerometer, **300**
electrical system block diagram, **218**
electrical systems, 216–17
electric field generation, **15**
electric field lines, 16
electric field strength, 10
electric torpedo, **293**
electrochemical horn fuzes, 324
electromagnetic (EM) spectrum, 100
electromagnetic energy, 5
electromagnetic energy modulation techniques, **44**
electromagnetic feed horns, **30**
electromagnetic fields, 37
electromagnetic fuze circuitry, **319**
electromagnetic horns, 30
electromagnetic impulse, 268–69
electromagnetic proximity fuzing, 318–20
electromagnetic pulse (EMP), 368
electromagnetic radiation, 27, 132
electromagnetic radiation theory, 54
electromagnetic rail gun launch package concept, **269**
electromagnetic wave field orientation, **16**
electromagnetic wave modes of propagation in the atmosphere, 22–25
critical frequency, 23
ground waves, 22
lowest usable frequency (LUF), 23
maximum usable frequency (MUF), 23
optimum working frequency (OWF), 23
radar line-of-sight, 24
radar line-of-sight equation solution, 25
sky waves, 22–23
space waves, 23–24
electromagnetic waves, 16–17
electronic attack (EA), 86–94
anti-ES countermeasures, 97
antiradiation weapons, 94
classes of, 87
deception jamming, 90–94
defined, 83
denial jamming, 87–90
effectiveness of, 97
effectiveness principles, 86–87
expendable decoys, 93–95
radar design philosophy, 97
radar parameter management, 95–96
signal-processing techniques, 96–97
target masking and modification (STEALTH), 94
transmitters, 87
electronic beam positioning, 21
electronic beam steering, 54
electronic combat (EC), 83
electronic combat (EC) equipment, 1
electronic combat (EC) receivers
direct detection receivers, 84–85
functions of, 83–84
superheterodyne receivers, 85–86
electronic countermeasures, 72
electronic protection (EP), 83, 94–98
electronic protection (EP) fixes, 96
electronic scanning and the phased array, 54–63
active electronically scanned array (AESA), 62–63
beam forming methods, 55–56
computation of required phase relationships, 59–62
energy reception, 59
frequency scanning, 57–58
other considerations, 63
phase scanning, 58–59
principles of operation, 54–55
time-delay scanning, 56
electronic support (ES), 83
electronic support (ES) equipment, 1
electronic support (ES) receiver, 84
electro-optical (EO) absorbers, 120
electro-optical (EO) and IR detectors, 133
electro-optical (EO) attenuation, 117–18
electro-optical (EO) detectors, **139**
electro-optical (EO) radiation sources, 108–9
electro-optical (EO) sensing
electro-optical (EO) attenuation, 117–18
electro-optical (EO) sources, 108–9
electro-optical (EO) spectrum, 100–102
plane geometry vs. solid geometry, 102
radiometric quantities, 103–8
radiometry vs. photometry, 102
reflected (external) radiation, 115–17
selective radiators, 114–15
spreading and divergence, 118–24
thermal radiation, 109–14
electro-optical (EO) signatures, **109**
electro-optical (EO) sources, 108–9
electro-optical (EO) spectrum, 100–102
electro-optical (EO) technology, 100
electro-optical applications, 320–21
electro-optics, **101**
electro-optic spectrum, **101**
electrostatic charge, 333
electrostriction, 193
electrostrictive effects, 192
electrostrictive transducers, 193
element pattern, 63
element spacing, 63
elements radiating in phase, **54**
elevation beam positioning logic, **61**
EMCON (RF emission control), 97
emission band, **129**
emissivity, 112–13
of common materials, **113**
defined, 112
endfire arrays, 21
endothermic reaction, 339
energy coupling, 345
energy reception, 59
energy transmissions, 26
engagement, 4
environmental problems, 168–69
equivalence of positive and negative phase shifts, **9**
E region, 23
erg, 372
erosion, 242
error, 227, 231–33
error correction, 66
error signal
derivative of, 233
input to integrating gyro, **251**
integral of, 233
ES antennas, **91**
escort jamming (self-screening), 88
evasion, 97
excessive tracking error, 92
exhaust gas management, 271
exhaust nozzle, 279
exhaust plume spectrum, **127**
exhaust temperature contours, **129**
exitance, **104**
expansion of continuous-rod warhead, 353
expendable bathythermograph (XBT), 160, **161**
expendable bathythermograph (XBT) probe, **161**
expendable decoys, 93–95
expendable floating decoys, **93**
expendable mobile target, **211**
expendable sound velocimeter (XSV), 161
explosion
beneath keel, **359**
near a submarine hull, **360**
explosion test, temperature for, 333
explosive power and power index, 340
explosive power testing, 333–34
explosive product decomposition reactions, **336**
explosive projectile test, 334
explosive propellant train, **276**
explosive reactions, 335–36
explosive reactions measurement
explosive power and power index, 340
explosive reactions, 335–36
heat of explosion, 338–40
oxygen balance (OB%), 337–38
volume of products of explosion, 337
explosives
defined, 330
heats of formation of common, **339**
impact sensitivity for common, **332**
and their uses for common, **332**
explosives categories, 330–32
detonation sequence and explosive train, 331–32
high explosives, 330–31
low explosives, 331

explosives characteristics, 332–35
availability and cost of, 332
brisance, 334
density, 334
hygroscopicity, 335
power, 333–34
sensitivity, 332–33
stability, 333
toxicity, 335
volatility, 334
explosive train, 276
exposure to sun's ultraviolet radiation, 333
exterior ballistics, 242–48
ballistics procedure, 247
Coriolis effect, 246–47
drag, 244–46
drift, 246
gravity, 242–44
prediction of future target position, 247–48
wind, 246
external combustion engine torpedo, **293**
extinction, 123
extreme IR (XIR), 101
eye (photopic) vs. night sky irradiation, spectral response of, 147

F/A-18 Hornet, 93
fallout, 369
false-alarm rates, 84, 86
false alarms, 179
false targets, 76, 90
Faraday's law of induction, 268
far IR (FIR), 101
fast Fourier transforms (FFT) method, 47
fast time constant (FTC), 36
fathometer, 209
Fat Man bomb, 364
feather, 207
feedback
closed-loop systems, 229–30
defined, 213
open-loop systems, 221–28
feedback loop, 229
feed horns, 30
ferrite phase shifter (phaser), **59**
fiber-optical cable, 305
field of view (FOV) error, 135–36
field orientation, 16
fields in space around a dipole antenna, 17
fields of view, **135**
field stop, 136
field strength, 12–13
figure of merit (FOM), 189–91, **190**
filter bandwidth, 47
final value theorem, 232
fin stabilization, 290
fireballs, 367
firebombs, 356
fire/burns, 377
fire control problem
defined, 237
described, 3
summary of the solution to, **261**
fire control solution, **248**
fire control symbols, 275
firing cutout mechanisms, 272
fish, 176
fission detonation, 364
fission process, **363**
fission vs. fusion, 362
fission warheads, 361, 362–63
5"/62 caliber gun, 264
5" projectile, **289**
fixed ammunition, 291
fixed objects imaging, 52
fixed targets, phase detector output for, **48**
fixed torpedo tubes, 267
fizzle, 364
flame holder, 281
flat plate halfpower point beamwidth, **194**
flechette round, 357
flexible connections used on pipes and hoses, **175**
flight path, 301
FLIR (forward looking infrared), 122
flow noise, 171–72
flux collection, **107**
flux lines, 16
fly ball governor, **213**
focal plane array (staring), 139
focal point of lens, **136**
focus depth, **137**
force multipliers, 93
form drag, 286
forward-looking infrared (FLIR) camera, 146
Fourier analysis principle, 14–15
fragmentation effects damage criteria, **352**
fragmentation round, **291**
fragmentation warheads, 350–53
fragment beamwidth, 351
fragmenting payload beamwidth, 351
fragmenting projectiles, 291
fragments distribution, **351**
fragment trajectories, 352
fragment velocity, 351
free gyro, 250
F region, 23
frequency
vs. absorption, **184**
described, 5–6
effects of changing, 63–64
vs. target size, 38
vs. time domain, 9
frequency agility, 50, 90, 95
frequency analysis, **14**, **202**
frequency analyzer, 202
frequency diversity, 90, 97
frequency-modulated (FM) radar altimeters, 45
frequency-modulated continuous wave (FMCW) systems, **45**
frequency-modulated CW radars, 44–45
frequency modulation (FM), **44**, **51**
frequency modulation chart, **44**
frequency scanning, 57–58
frequency-scanning radar, **58**
frequency spectrum, **5**
friction, 333
friction drag, 286
friction pendulum test results, **333**
fuel, merit or quality of, 278
fuel-air-explosive (FAE) vs. chemical explosives, 356
fuel-air-explosive (FAE) warheads, 348, 356
fuel-air-explosive (FAE) weapons, 318
fully stabilized tracking subsystems, 68
functional reliability, 314, 326–27
fusion of deuterium with tritium, 362
fusion warhead design, 365
fusion warheads, 361, 364–65
future target position, 238
fuze assembly, 288
fuze classification, **318**
fuze defined, 314
fuze system, **314**, **327**, **329**
fuze system classification, 317
fuze system functions, 314–17
safing and arming (S&A) devices, 314–16
target detection and recognition, 316–17
warhead initiation, 317
fuzing reliability, 326–27
functional reliability, 326–27
safety reliability, 327

gain, 34, 233
galvanic fuzes, 324
galvanic-type mine fuze system, **326**
gamma radiation, 364, 368, 371, 375
gamma-ray doses, slant ranges, **381**, **382**
gamma-ray reflector, 365
gate, 77
gate dimensions, 77
Geiger counter, 372
generation of an electric field about dipole antenna, **15**
generation of electromagnetic waves, 15–16
generator principle, 321
Geneva escapement, 315
Gen 0 M3 infrared sniperscope, **147**
Gen I I2 electrostatic tube technology, **148**
Gen I I2 tube technology, **148**
Gen II AN/PVS-5, **149**
Gen II I2 tube technology, **149**
Gen III ANVIS-6, **150**
Gen III FLIR systems, 139
Gen III I2 tube technology, **150**
geometric line-of-sight, 24–25
geometricoptics, 121, 122
geothermic chemical reactions, 335
German V-2 rocket, 308
global positioning system (GPS), 298, 311
global positioning system (GPS) receivers, 312
global positioning system (GPS) segments, **312**
gradiometer, 321
grating lobes, 63
gravity, 242–44
gravity and parasitic drag, **245**
gravity launchers, 272
gravity-type propulsion, 282
gray, 373
graybodies
defined, 112
spectral exitance of, **112**
Stefan-Boltzman law modification for, 112
and temperatures, **113**
grazing angle, 163
ground penetrating fuzes, 325
ground state, 132
ground wave propagation, **22**

ground wave radar, 64
ground waves, 22
ground zero, 368
guidance phases of missile flight, **301**
guidance systems
  accelerometers, 299–300
  boost (initial) phase, 300–301
  fundamentals of, 298–99
  midcourse phase, 301
  sensors, 298–99
  terminal phase, 301
guidance systems types, 304–12
  control guidance, 304–6
  homing guidance, 306–8
  self-contained guidance, 308–12
guided missiles
  accelerometers in, **300**
  control surface configuration, 287
  functional systems of, **288**
  zero-length launchers and, 270
  see also missile architecture
guided missiles, specific types
  AGM-88 high speed antiradiation missile (HARM), 94
  air launched cruise missiles (ALCMs), 2, 281
  Basic Point Missile Defense System, 66
  cruise missiles, **130**
  Harpoon missiles, 281
  home-on-jam (HOJ) type missile systems, 88
  ICBM (intercontinental ballistic missile), 309
  Israeli Python missiles, 138
  Javelin missiles, 271
  Nike-Ajax missile systems, 71
  Polaris missile, 308
  Sidewinder missiles, 138
  SLBM (submarine launched ballistic missile), 309
  Standard missile, 271
  Stinger missiles, 271
  Tomahawk missiles, 281
guided projectiles, 291
gun ammunition architecture, 289–92
  aerodynamic drag, 289–90
  aerodynamic stabilization, 290
  fragmenting projectiles, 291
  guided projectiles, 291
  gun propellant configuration, 291
  penetrating projectiles, 290–91
  projectile parts, 289
  special cases, 292
  special purpose projectiles, 291
gun barrel cross section, **263**
gun barrels, 263–64
gun bore erosion, 242
gun design warhead, 360
gun launchers
  for aircraft, 266
  catapult ramming of field ammunition in, **274**
gun mount overhead view, **216**
gun mount system, **216**
gun orders, 275
gunpowders, 240
gun projectiles
  aerodynamic forces on, **290**
  details, **289**
  for RF proximity fuze for a gun projectile, **319**
  see also gun ammunition architecture
gun propellant configurations, **291**
gun recoil system model, **214**
guns, 262
gun strength, 263
gun strength curve, 239
gun temperature, 242
gun tube length, 241
gun-type fission warhead "Little Boy," **363**
gun-type launchers, 263–64
Gurney constants, 334, **351**
gyro drift, 251
gyro precession, **250**
gyroscopes (gyros), 68, 250–53
gyroscopic effects, 246
gyroscopic inertia, 250

half-channel propagation, **166**
half-power points, 19, 33
half-wave dipole antenna, **18–19**
hard target penetrator warheads, 357
harmonization, 260
Harpoon canisters, **271**
Harpoon missiles, 281
hazards to ordnance (HERO), 41
hazards to personnel (RADHAZ), 41
head-up displays (HUDs), 260
health effects, **374**
heat/heating, 37, 333
heat of explosion
  described, 338–40
  for nitroglycerin, 340
  for TNT, 340
  see also Gurney constants
heat of reaction, 340
heats of formation, 338, **339**
heat sources on armored vehicles, **131**
height of burst (HOB), 378, 379
high-altitude burst, 369
higher-energy state, 115
high explosives, 330–31
high-explosive train, 331
high-explosive train sequence, **331**
high frequency (HF), 22–23
high-frequency radars, 63
high-resolution radars, 51–53
  inverse synthetic aperture radar (ISAR), 53
  pulse compression, 51
  synthetic aperture radar (SAR), 51–53
high-scan-rate search radar, 75
high-speed low-flying aircraft, 92
high-strength, extra-long guns, 292
Hiroshima, 363
histogram, 342–43
hollow charge effect, 353
home-on-jam (HOJ) type missile systems, 88
home-on-jam (HOJ) weapons, 88
homing accuracy, 307
homing guidance, 306–8
homing guidance methods, **306**
homing logic, 301–4
Hooke's law, 227
horizontal component of peak dynamic pressure, **379**
horizontal polarization, **17**, 18
howitzers, 262
howler principle, 161
hydrodynamic noise, 175
hydrodynamic pressure, 354
hydrogen cyanide, 386
hydrophone array, 201
hydrophones, 156, 194, 321
hydrostatic (pressure) fuzing, 323
hygroscopicity, 335
hypersonic cruise missile engine, **282**
hypersonic ramjet, **281**

ICBMs (intercontinental ballistic missiles), 309
illuminating projectiles, **356**
illumination warheads, 356
image intensifier tubes, **151**
imaging of fixed objects, 52
impact mine fuzes, **326**
impact sensing devices, 316
impact sensitivity, **332**
impact testing, 332
implosion warhead design, **364**
implosive warheads, 364
impulse launch, 262
impulse launchers, 262–69
  ejector launchers, 267–68
  electromagnetic impluse, 268–69
  gun-type launchers, 263–64
  muzzle brake, 265–66
  recoil/counterrecoil systems, 264–65
  recoilless systems, 266–67
  soft recoil systems, 266
impulse propulsion, 276
impulse vs. scaled distance, **334**
incendiary bombs, 356
incident wave, 347
incremental phase shift, 59
index of refraction, 6
indices of refraction for common substances, 6
indoor radon, 376
induced drag (ID), 284, **285–86**
induced drag curves, 286
induced field, 207
inductor mechanisms, 321
inertial guidance, 308–9
inertial navigation guidance system, 299
infinite diameter detonation velocity, 334
influence sensing, 317
infrared CCDs (IRCCDs), 142
infrared detection, 207
infrared radiation (IR), 99, **101**
infrared sources, 108
infrared wavelengths, 100
initial and final states principle, 339
initial flight orientation, 275
initial radiation, 368, 371
initial static velocity, 350
initial velocity, 241, 351
instantaneous field of view (IFOV), 136
instantaneous frequency measurement (IFM) receiver, 85
intake duct, 280
integrated automatic detection and tracking (IADT), 82
integrating gyro, 251

integration period, 318
intensity, **105**, 155–56, 186
intercept receiver, 210
interference, 12–14, **13**
interference drag, 286
interferometer (phase comparison monopulse), 307
interior ballistics
  described, 238–42
  and impulse propulsion, 276
  muzzle velocity, 241–42
  propellant types, 240–41
  rifling, 241
interior ballistics diagram, **239**
intermediate frequency (IF), 85
intermediate high explosives, 331
International Commission on Radiation Protection (ICRP), 375
intrinsic detectors, 140
inverse conical scan, 69–71
inverse conical scan (inverse gain) methods, 92
inverse first power spreading, 182
inverse Laplace transform, 217
inverse square law (spreading), **108**
inverse square law of radiation (spreading), 107–8
inverse synthetic aperture radar (ISAR), 53
ionization, 370
ionizing radiation, 372, **374**
ionizing radiation effects, 370–71
ionosphere, 22, 23
ionospheric regions, **23**
ionospheric scatter, 24
ionospheric scatter propagation, **24**
IR detection diagram, **126**
irradiance, **104**, 107
IR signatures, **127**, **128**, 130
IR spotlight, 147
isotropic antennas, 17
isotropic blast patterns, **345**
isotropic propagation, 345
isotropic radiators, 104
isovelocity, 164
isovelocity condition, **164**
Israeli Python missiles, 138
iterative procedure, **248**

jammer signal, **71**
jammer spectral power density, **89**
jamming, **87**, **88**
jamming tactics, **89**
Javelin missiles, 271
JDAM, **341**
jet precision, 355
jet-propulsion classification, 277
jet tip velocity, 354
Joint Munitions Effectiveness Manuals (*JMEMs*), 341

Kalman filter, 80
"Kentucky windage," 66
kinetic energy equation, 352
kinetic energy warhead, 355
Kirchoff's voltage law, 216

lagging a target, 80
Lambertian sources, 108
Lambertian surface, 106–7, **107**
Lambert's cosine law, 106
Lame's law, 263
laminar flow, 171
laminar vs. turbulent flow, **171**
Laplace transform pairs, **217**
Laplace transforms, 217
large range resolution, **31**
laser mediums, **133**
lasers (light amplification by stimulated emission of radiation)
  characteristics of, 133–34
  defined, 132
  for detection of shallow underwater objects, 207–8
  in electro-optical (EO) fuzing applications, 320–21
  in electro-optical (EO) sensing, 100, 108
  principles of operation, 132–33
  quantum efficiency of, 144
  turbulence and, 122–23
  see also ring laser gyro (RLG)
laser TDD, **321**
late range gates, 75
launcher cutouts, 272
launching system, 262
launching systems functions, 272–76
  control, 274–75
  loading equipment, 274
  storage, 272–73
  summary, 275–76
  transfer equipment, 273
layers, deep-sea, **162**
LD 50/30, 374
leading a target, 80
leaflet bomb, **357**
leaks, 34
left-hand circular polarization (LHCP), 18
lens, 134, **136**
lethal burst interval, 319
lethal kinetic energy, 352
leukopenia, 373, 375
Lewisite, 387
lifetime of higher-energy state, 114
lift characteristics, **284**
lift coefficient, 284
lightweight projectiles, 292
linear antenna arrays, 20–21
linear array directionality, **204**
linear polarization, 18
line-of-fire (LOF), 274, 301
line-of-position (LOP)
  on the earth's surface, **311**
  fix from two, **311**
  with GPS, 311
line-of-sight (LOS)
  collimation and, 259
  and ducting, 24
  error correction to, 66
  and lead angle, 275
  minimum detectable speed in, 50
  relative speed in, **43**
  and the tracking line, relationship between, **67**
  see also radar line-of-sight
line-of-sight (LOS) path, 303–4, **304**
line source, 390
liquid-damped system, **300**
liquid propellants, 279
Little Boy bomb, 363
loading density, 334
loading equipment, 274
lobe on receiver only (LORO), 71
lobe on receiver only (LORO) radar, 71
lobes, 34
lobe-switching, 69
local fallout, 369
localization, 1
local oscillator (LO), 85
lockers, 273
longitudinal wave, 154, **155**
long-range land attack projectile, **292**
long-range weapons, 269
longwave IR (LWIR), 101, 118
look angle, 307
lower frequency limit, 205
lowest usable frequency (LUF), 23
low explosives, 331
low-flying aircraft, 64
low frequency (LF), 22–23
low probability of intercept (LPI), 97
low-side-lobe levels, 96
low-supersonic ramjet, **281**
lunar illumination, 115

M-60 tank signature, **131**
Mach effect, 347
machinery noise, 171
mach number vs. drag coefficient, **290**
Mach stem, **347**
Mach wave, 347
magazines, 272–73
magic tees, 73
magnetic anomaly detection (MAD), 321
magnetic anomaly detection (MAD) sweep, **209**
magnetic anomaly detectors (MAD) systems, 153, 208–9
magnetic field generation, **16**
magnetic fields, **208**
magnetic signature, 208
magnetostatic fuze applications, 321
magnetostrictive effects, 192, 193, **194**
magnetostrictive transducers, 193–94
magnetron, 28
magnification, 134–35
Magnus effect, **246**
main (bursting) charge, 331
marine fouling, 172
marine mammals, 177
marker warheads, 357
Masker belts, **174**
master control station (MCS), 312
maximum radiation, 366
maximum unambiguous range, **50**
maximum unambiguous speeds, 49
maximum unambiguous velocities, **50**
maximum usable frequency (MUF), 23

Maxwell's theory, 15
MDS (minimum discernable signal), 35
mean impact point (MIP), 344
measurement of radiation
  see radiation measurement
mechanical systems, 215
Mechan Range Unit, 74
medium frequency (MF), 22–23
merit of fuel, 278
mesosphere, 118
metal oxide structure (MOS), **143**
metal-oxide-structure (MOS) devices, 142
microbar, 156
microchannel plate (MCP), 149
micropascal, 156
midcourse phase, 301
middlewave IR (MIR or MWIR), 101, 118
Mie scattering, 121, 122
millimeter wave radar, 63
mine actuation distance, **324**
mine case, 295
mine classification, 294–95
mine components, 295–96
mine contact fuzes, 324
minefields, 294
mine performance data, 323–24
mine positioning, **295**
mine warheads, 359–60
minimum detectable speed, 50
minimum discernable signal (MDS), 35–36, 144
minimum lethal kinetic energy, 352
minimum range, 31
minimum signal for detection, 35
minimum velocity limits, 50
miss distance, 342
miss distance plot, **344**
missile architecture, 282–89
  autopilot, 288
  control surface configuration, 287–88
  control systems, 289
  drag, 284–87
  guidance system, 288
  lift, 283–84
  missile components, 288
  propulsion, 288
  stability, 287
  warhead, 288
missile flight guidance phases, **301**
missile force diagram, **283**
missile orders, 275
missiles, 298
  see also guided missiles
missile signatures, 131
mixer, 42
MK-32 torpedo tubes, **267**
MK-46 torpedo, 295
mobile mines, 295
modulation, 44
modulator, 28
monochromatic lasers, 134
monopropellant fuels, 293
monopropellants, 279
monopulse, 71–73
monopulse radar signal comparator circuit, 73
monopulse radar system, **71**
monopulse trackers, 73
monotrack radar, 92
monotrack systems, 86
moon light, 115–16
moored mines, 295, **296**
mortars, 262
motion detectors, 49
movement detectors, 49
moving target indicator (MTI) radar, 46
moving target phase detector output, **48**
MTI mode, 48–49
MTI phase comparator radar system, **48**
MTI processor cancellation circuit, **48**
MTI signal processing, 95
multiple frequencies combination, **14**
multiple hop propagation, 64
multiple paths, 168
multiple propagation paths, 63–64
multiple vertical beams, **201**
multipurpose fragmentation round, **291**
multistatic sonar, 206
Munroe effect, 353
mushroom clouds, 367, **368**
mustard-lewisite mixture, 387
mustards, 386–87
muzzle brake, **265–66**
muzzle velocity, 239, 241–42
MWIR plume image, **131**

Nagasaki, 364
napalm, 356
narrowband display, 203
narrowband grams, **203**
narrowband sources, 170, **171**, 180
National Council on Radiation Protection and Measurement (NCRP), 375
natural illumination, 115–17
naval mines architecture, 294–96
  classification, 294–95
  mine components, 295–96
navigational guidance systems, 308
near IR (NIR), 101
negative durations, 346
negative feedback, **227**
negative temperature gradient, 164
negative velocity gradient, **165**
nerve agent poisoning symptoms, 385
nerve agents, 385–86
neutral burning grains, **240**
neutralization, 3, 4
neutron doses slant ranges, **382**
neutron leakage, 364
neutron life cycle, 363
neutron radiation, 368, 371–72
neutron reflector, 364
Newton's third law of motion, 262
night sky illumination, 116
night sky irradiation
  vs. eye (photopic), spectral response of, **147**
  vs. NVGs, spectral response of, **147**
night vision devices (NVDs), 146–51
  first-generation image intensifiers, 147–49
  fourth-generation image intensifiers, 150–51
  generation 0 image converters, 147
  image intensifier (I2) tube, 146–47
  second-generation image intensifiers, 149–50
  third-generation image intensifiers, 150
night vision goggles (NVGs), 115, 116, 122, **147**
night vision image intensification, **147**
night vision imaging system (NVIS), 146
Nike-Ajax missile systems, 71
nitroglycerin, 336, 337, 338, 340
noise, 123, 179, 187
noise-distorted radar pulse, **31**
noise-equivalent-power (NEP), 35, 145
noise jamming, 87
noise level (NL), 157, 186
noise limited situations, 188
noise limited systems, 145
nonacoustic detection, 207–9
  infrared detection, 207
  laser, 207–8
  magnetic anomaly detectors (MAD) systems, 208–9
  radar, 207
  visual, 207
non-antenna coupling, 41
noncoherent energy, 6
noncoherent waves, 32
nonisotropic blast patterns, **345**
nonisotropic propagation, 345
nonpersistent chemical agents, 385
nonproximity-type fuzes, 324–26
  command fuze, 325–26
  delay fuzes, 325
  point detonating fuzes, 324–25
  time fuzes, 324
normal distribution, **343**
normal-mode theory, 162
noticeable reaction, 332
nozzle components, **279**
nuclear, biological, and chemical (NBC) weapons, 361
nuclear binding energy, 363–64
nuclear blast effects prediction, 377–82
  radiation dose, 381–82
  using scaling laws, 377–80
  thermal radiation, 380–81
nuclear explosions, **366**
  damage data for various targets, **377**
  lethal ranges to exposed personnel, **369**
  mushroom cloud from, **368**
nuclear explosions characteristics, 366–72
  additional effects of, 368–69
  blast wave, 366–67
  burst types, 368–69
  ionizing radiation effects, 370–71
  nuclear radiation, 368
  radiation effects, 371–72
  thermal pulse, 368
nuclear radiation, 368
nuclear weapons, 361–66
  fission vs. fusion, 362
  fission warheads, 362–63
  fusion warheads, 364–65
  gun design warhead, 360
  implosive warheads, 364
  principles of operation of, 361–62
  proliferation of, 365
  safety of, 365–66

nuclear weapons damage prediction, 376–77
null-to-null bandwidth and pulse width, 37
nutating feed horn, 69

object and image distances, 136
oceanic fronts, 159
ocean propagation, 123
ocean traffic, 176
ocean transmission windows, 123
offset feed and beam direction changes, 72
ogival curve, 289
ogive, 289
omnidirectional noise, 185
one-fourth power relationship, 40
105-millimeter light-towed howitzer, 266
120-millimeter kinetic energy antitank projectile, 355
open-loop speed control, 227
open-loop systems, 221–28, 228
open-loop transfer function (OLTF), 229–30
optical characteristics of EO and IR detectors, 133
optical line-of-sight (LOS) reference, 259
optical region, 38
optical sensors, origins and development of, 99
optics, 134–37
  depth of focus, 136–37
  field of view (FOV) error, 135–36
  magnification, 134–35
  stadimeter ranging, 137
optimum working frequency (OWF), 23
order of priorities for explosive product decomposition reactions, 336
oscillating mines, 295
oscillator, 42
out-of-battery firing, 266
overdamped response, 20
overdamped time response, 220
overlapping return signals, 31
overpressure, dynamic pressure, and wind velocity, 376
overshoot, 221
over-the-horizon radar (OTH), 63–64
oxygen balance (OB%), 337–38

pair production, 371
palliative dose, 374
parabolic antennas, 19–20
parabolic cylinder radiation pattern, 20
parabolic reflector, 20
parallax, 258
parallax correction, 258
parallel scanners, 139
parallel scanning, 139
parametric data, 84
parasitic curves, 286
parasitic drag, 245, 284, 286
parasitic elements, 21
parasitic reflector radiation patterns, 21
partially stabilized tracking subsystems, 68
partial reactions, 335
particles scattering, 121
passive acoustics, 97
passive acoustic torpedoes, 322
passive beamforming, 202
passive homing, 293, 307
passive hydrophone array, 201
passive IR, 97
passive sensors, 153
passive sonar equation, 186–87
passive sonar system, 200
passive source level (SL), 180
passive systems, 200–203
  beamforming processor, 201–2
  broadband display, 202
  frequency analyzer, 202
  hydrophone array, 201
  narrowband display, 203
pathogens, 389, 390
Patriot missile radar, 61
peak dynamic pressure curve, 379
peak overpressure
  high-pressure region, 379
  intermediate-pressure region, 378
  low-pressure region, 378
  vs. shock (blast) wave, 367
  as shock wave, 367
  shock wave development of, 346
  vs. time relationship of blast wave, 346
  for TNT, 349
peak overpressure curves, 380
peak power, 27
peak time, 221
peak underwater pressure, 358
penetrating projectiles, 290–91
penetration depth, 354
percent overshoot, 221
performance degradation, 384
permanent magnetic flux lines, 208
permanent thermocline, 162
persistent chemical agents, 385
phase, 8–9
phase comparator, 48
phased-array radar antenna section, 59
phased-array techniques, 60
phase difference
  due to distance from source, 12
  due to time difference, 13
  between two waves, 8
phase front distortion (crosseye), 92
phase lag, 58
phaser, 59
phase scanning, 58–59, 59
phase shift, 8–9, 10, 11, 252–53
phase shift, incremental, 59
phases of guidance, 300–301
phenyldichorarsine, 387
phosgene, 386
phosgene coating, 387
photoconductors, 140
photoelectric effect, 371
photoemissive detection, 141
photometric energy, reflected, 116
photometric quantities, 103
photometric units, 102
photometry vs. radiometry, 102
photomultiplier, 141
photomultiplier tube (PMT), 141
photon absorption, 119
photon detectors, 140
photons, 109
photopermissive devices, 141
photovoltaic detectors, 140–41
photovoltaic energy diagram, 140
piezoelectric effects, 192, 193
Pillenwerfer, 210
pitch and roll, 257
Planck's constant, 109
Planck's law, 110, 114
plane geometry vs. solid geometry, 102
plane wave, 10
plane wave spreading, 10
plan position indicator (PPI), 29, 198
plastic armor defeat charge, 355
plate dent test, 334
platform shape alteration, 94
plume calculations, 129
plume radiance, 130
point detonating fuzes, 316, 324–25
point of detonation, 319
point sources, 34, 108
point sources vs. line source, 391
point source vs. extended source, 104–5
Polaris missile, 308
polarization, 18
position controls, 213
position error magnitude and direction, 69
position feedback, 213
positioning feedback, 73–74
positioning logic, 61
positive durations, 346
positive feedback loop, 229
positive phase of shock (blast) wave, 346
positive temperature gradient, 164
positive velocity gradient, 165
post-detection pulse compression (PDPC), 51
powder temperature, 241
power, 95, 107, 333–34
power density, 181
power density equations, 39
power driving devices, 74
power gain, 34
power gain vs. directive gain, 35
power index
  of deuterium with tritium, 362
  and explosive power, 340
  in weapon damage calculations, 349
power supplies, 28
PPI display, 29, 198
PPS (precise positioning system), 312
Prairie air system, 174
Prairie Masker, 173, 174, 212
precession, 250
precursors, 385
predicted intercept point (PIP), 2
predicted intercept point (PIP) display, 47
prediction
  of nuclear blast effects. *See* nuclear blast effects prediction
  of nuclear weapons damage. *See* nuclear weapons damage prediction
prediction methodology of radiation damage, 377
prediction of future target position, 247–48
preset guidance method, 308
preset homing logic, 301–2

preset paths, 301
pressure
  defined, 154, 156
  vs. scaled distance, 334
  and sound speed, 159
pressure (profile) drag, 286
pressure signature of ships, 323
pressure strength curve, 263
pressure travel curve, 239, **241**
pressure variation due to ship travel, **323**
pressure vs. time for underwater explosion, **358**
primary (trigger), 365
primary magazines, 273
primer, 331
principles of operation
  of electron beam steering, 54–55
  of lasers, 132–33
  of nuclear warheads, 361–62
  of radar, 26–30
prism dispersion of white light, **6**
probabilities of occurrence, 327–28
probabilities of success, 329
probability of damage
  described, 341–42
  for GPS-guided 500-pound bomb vs. mobile missile launcher, **342**
  for unguided 500-pound bomb vs. mobile missile launcher, **342**
  for unguided 500-pound bomb vs. vehicle, **342**
probability of intercept (POI), **84**, 85
probability theory, 327–29
  mututally exclusive events, 327–28
  parallel systems, 328–29
  series systems, 328
  statistically independent events, 327
programmed preset, 302
progressive burning grains, **240**
projectile impact fuze, **325**
projectile parts, 289
projectile weight, 241
projector, 192
prompt radiation, 368
propagating wave, 7
propagation, 345
propagation in liquids and solids, 124
propagation loss curves, **189**, **190**
propagation losses, 118
propagation paths
  bottom bounce, 168
  convergence zone, 167–68
  direct-path propagation, 166
  environmental problems, 168–69
  half-channel-propagation, 166
  multiple paths, 168
  of sky wave radar, 63
  sound channel propagation, 166–67
  surface-duct propagation, 166
  thermal structure, 164–66
propagation speed, 154–55
propagation windows, 63
propellant energy, **241**
propellants, 277–79
propellant types, 240–41
propelled warhead mines, 295
proportional, 304
proportional control, 233
proportional-derivative (PD) control, 233
proportional-integral (PI) control, 233–34
proportional-integral-derivative (PID) control, 234–35
proportionality factor, 156
proportional navigation, 304
proportional navigation flight path, **304**
propulsion subsystem subelements, 279
propulsion systems, 293
protection standards, 375–76
proximity fuzes
  combination fuzing, 323
  electromagnetic proximity fuzing, 318–20
  electro-optical applications, 320–21
  FMCW systems in, 45
  hydrostatic (pressure) fuzing, 323
  magnetostatic fuze applications, 321
  mine performance data, 323–24
  in missile architecture, **288**
  seismic fuzing, 323
  underwater electric potential (UEP), 323
proximity of warhead to target, 341
proximity sensing device, 317
psychological warheads, 357
pulse compression, **51**
pulse compression radar, 95
pulse compression ratio (PCR), 51
pulse-Doppler (PD) deception, 92
pulse-Doppler (PD) radar, 46, **50**
pulse-Doppler (PD) radar mode, 46–48
pulse-Doppler (PD) radar system, **47**
pulse-echo-type radar, 28
pulse integration technique, 36
pulse length, 96
pulse-modulated Doppler system, **46**
pulse repetition frequency (PRF), 27, 31–32, 46, 95–96, 161
pulse repetition rates, 50
pulse repetition time (PRT), 26, 27, 31, 95
pulse shape and width, 30–31
pulse transmission, 26–27, **27**
pulse width (PW)
  and bandwidth, 50
  vs. central frequency bandwidth, **37**
  minimum detection range and, 27, 31
  and null-to-null bandwidth, 37
pumping system, 133
pursuit paths, **303**
pyroelectric detectors, 142
pyroelectric effect, 142
pyrotechnic warheads, 356–57
Pythagorean Theorem, 380

quality factor, **373**
quality of fuel, 278
quanta, 109
quasi-optical reflectors, 19–20
quenching, 196
quenching power limit, 196
quieting, 173

radar, 26, 207
radar-absorbent material (RAM), 94
radar altimeter, **45**, 310
radar blind speeds, 96
  see also blind speeds
radar cross section (RCS), 37–39, **38**, 94
radar design philosophy, 97
radar energy, 5
radar environmental considerations, 40, 41
radar frequency band designations (historical), **32**
radar frequency bands, **38**
radar horizon, **24**
radar line-of-sight, 24
radar line-of-sight equation solution, 25
radar line-of-sight geometry, **25**
radar line-of-sight horizon, 24
radar mechanization, 28–30
  antennas, 29
  display, 29
  duplexers, 29
  power supplies, 28
  receivers, 28
  synchronizers, 28
  transmitters, 28
  wave guide, 29–30
radar parameter management, 95–96
radar ppi scope, **88**
radar principles of operation, 26–28
  pulse transmission, 26–27
  range determination, 27–28
radar range equation, simplified, 39–41
radar range formula, 31
radar range resolution, 31
radar range-time relationship, **28**
radar system components, **28**
radar warning receivers (RWRs), 85
radial velocity, 42–43
radiance, 105–6
radians, 9, 102
radiant energy, 103
radiant exposures, **381**
radiant flux (power), 103–4
radiant intensity, **105**
radiation absorbed dose (rad), 372, 373
radiation biological damage limits, 381
radiation cross section
  half-wave dipole antenna, **18**
  for half-wave dipole antenna, 19
radiation dose prediction, 381–82
radiation effects, 371–72
radiation exitance, 104
radiation exposure, 377
radiation field, 18
radiation injuries, 374
radiation limits, **375**
radiation measurement, 372–76
  protection standards, 375–76
  somatic effects of, 373–75
  units of measure, 373
radiation path length difference, **55**
radiation pattern
  of beam, **19**
  of parabolic cylinder, **20**
  for two-element array, **21**
radiation sickness phases, 375
radiation sources, **127**
radiation warheads, 356

radioactive contamination, 368
radio frequency (RF), 27
radiometric quantities, 103–8
  flux collection, 107
  inverse square law of radiation (spreading), 107–8
  irradiance, 104
  Lambertian surface, 106–7
  point source vs. extended source, 104–5
  radiance, 105–6
  radiant energy, 103
  radiant flux (power), 103–4
  radiant intensity, 105
  radiation exitance, 104
  source geometry, 108
radiometric units, **103**
radiometry, 102
radius of curvature, **164**
radon, 375–76
rail gun, 268
rail gun design, **268**
rail launchers, 269–70
rail-spade ramming, 274
rail travel, 270
rain, 37
ramjet engines, 281–82
ramming cycle, 274
random error, 344
random search, 82
range
  calculated as, 45
  vs. Doppler shift, showing positive and negative images, **53**
range-cutoff of TDD sensor, 318
range deception, 91–92
range determination, 27–28
range error, **75**
range-gate circuit, 318
range-gate mode, 91
range gates, **75**
range gating system, 75
range rate, 26
range resolution and pulse compression, 51
range resolution formula, 51
ranges at different FOMs based on propagation loss curve, **190**
ranges of frequency-modulated continuous wave (FMCW) systems, 45
range tracking, 74–75
range wind, 246
rarefaction, 367
rate gyro, 251
ray, 163
Rayleigh region, 38
Rayleigh scattering, 121, 122
ray theory, 162–64, **163**
RBOC (rapid blooming offboard countermeasures), 93
reaction launchers, 269–72
  canister launchers, 270–72
  chain ramming of guided missile in, **274**
  launcher cutouts, 272
  rail launchers, 269–70
reaction motor, 277
reaction propulsion, 276–82
  advantages by type, **283**
  described, 276–77
  disadvantages by type, **283**
  jet-propulsion classification, 277
  propellants, 277–79
  propulsion subsystem subelements, 279
  ramjet engines, 281–82
  turbojet engines, 280–81
ready service lockers, 273
receivers, 28, 35–36, 198
receiver sensitivity, 84
reciprocal bearing, 43
recoil, 264
recoil/counterrecoil, **265**
recoil/counterrecoil systems, 264–65
recoilless rifle, **267**
recoilless systems, 266–67
recording system, 84
rectangular coordinates, 79
rectangular coordinate systems, **254**
rectangular coordinate systems vs. spherical coordinate systems, **255**
redundancy, 327
reference element, 258
reference element and sign convention, **60**
reference frame rotation (coordinate transformation), 255–57
reference frames and coordinate systems, 248–50
reference frames and coordinate systems, relationship between, **256**
reference frame stability, 249
reference frame translation, 257–58
reference gyroscope, 68
reference planes, 248
reference signal, 42
reflected (external) radiation, 115–17
  artificial illumination, 117
  natural illumination, 115–17
  reflectivity (albedo), 117
reflected photometric energy, **115**
reflection
  angles of, **10**
  phase shift on, **11**
reflection coefficient, 10
reflection of the blast wave from a surface and Mach stem, **347**
reflectivity (albedo)
  of common materials, **113**
  defined, 111
  of terrain/objects, 116, **117**
reflectors, 93
reflectors, neutron, 364
reflector shapes, **19**
refraction, 10, **11**, **24**
regions of the ionosphere, **23**
rejection of stationary targets, **46**
relative bearings, 29, 43
relative bubble energy (RBE), 334
relative motion, 237–38
relative speed, **43**
relative wind, 283
relay, 332
reliability of fuzing, 326–27
rem (roentgen equivalent to man), 373
repeater jammers, 90
repeaters, 90
residual radiation, 368, 371
resilient mounts, **175**
resolution or localization, 3
resonance region, 38
responsivity, 143
responsivity curve, **143**
rest time (RT), 27
retransmission homing, 307
returning "skin" echo, **71**
return pulse amplitude and target position, **70**
return strength, **67**
reverberation, 184
reverberation, transmitted pulse, and echoes, **199**
reverberation level, 186
reverberation limited (RL) situations, 188
reverberation-limited active sonar equation, 188
RF emission control (EMCON), 97
RF proximity fuze for a gun projectile, **319**
rifling, 241, 290
right-hand circular polarization (RHCP), 18
right-hand rule, 15
rigidity in space, 250
ring laser, 252
ring laser gyro (RLG), **251**, 251–53
rocket assist (RAP) motor, 291
rocket assisted projectiles, 292
roentgen, 372
roller-path planes, 258
root locus form, 218
root mean square (rms) of pressure, 155
rotating band, 289
rotational system block diagram, **218**
rotational systems, **214**, 215, 216
rotation rate, 252–53
"Rubber Duck," 93
rules of engagement (ROE), 2

safety of nuclear weapons, 365–66
safety reliability, 314, 327
safing and arming (S&A) device, **316**, 331
Sagnac effect, 252
sailboat mast moving through an arc length, **53**
salinity, 159
sand test, 334
sarin, 385
satellite radar systems, 52
scaled distance
  vs. impulse, 334
  vs. pressure, 334
scaling law, 349, 366
scan frequency, 90
scanners, 138–39
scanners and detectors, 137–46
  bolometers, 142
  charge-coupled devices, 142–43
  detector responsibility, 143
  detectors, 139–43
  focal plane array (staring), 139
  parallel scanners, 139
  photoconductors, 140
  photon detectors, 140

photopermissive devices, 141
photovoltaic detectors, 140–41
pyroelectric detectors, 142
scanners, 138–39
sensitivity, 144
sensor bandwidth, 144
sensor response, 144
serial scanners, 138–39
target detection, 144–45
thermal detectors, 141–42
scanning in elevation, **60**
scan patterns, **61**, 96
scan rate, 34
scar, 207
scattering, 15, **121–22**
scattering coefficients, **122**
scattering losses, 183–85
scattering patterns, 121
scatter propagation, **24**
scintillations, 122
scramjet (supersonic combustion ramjet), 282
s domain, 217
sea-based sources, 132
search coil, 321
seasonal thermocline, 162
sea states, 53
sea states and Beaufort Wind Scale, **176**
seismic fuzing, 323
seismic noise, 176
selectively aimable warhead (SAW), 352
selective radiators, 110, **114–15**
selective radiators output, **110**
selective source, 123
self-contained guidance, 308–12
self-noise, 170–71, **175**
self-noise levels vs. speed, **172**
self-screening (escort jamming), 88
semiactive homing, 306
semifixed ammunition, 274
semifixed propellant charges, 291
sensitivity, 35, 144, 332–33
sensitivity time control (STC), 36
sensor bandwidth, 144
sensor mechanisms, 316
sensor response, 144
sensors, 298–99
separation of targets, Doppler frequency gates for, **46**
sequential lobing, **69**
serial-parallel processing, 138
serial scanners/scanning, **138–39**
serpentine feed assembly, **58**
serpentine wave guide line, 57
servomechanisms, **67**, 214
servo systems, 66–75
azimuth tracking, 69–73
described, 66–68
position error magnitude and direction, 69
positioning feedback, 73–74
power driving devices, 74
range tracking, 74–75
stabilization, 68
velocity feedback, 74
servo tracking systems, **67**
settling time, 221
76 mm naval gun, **264**
shading, 194
shadow zone, 164, 165
shallow water regions, 169
shaped charge definitions, **354**
shaped charge effect, 353
shaped charge sequence, **354**
shaped charge warheads, 353–55, **354**
shape drag, 244
shielding rule, 371
ships
detection of, 64
pressure signature of, 323
ship's hull and earth's magnetic field, **208**
ship signatures, **132**
ship travel
pressure variation due to, **323**
sound variation due to, **322**
shock blast wave
attenuation of, 345
dynamic overpressure as, **367**
during explosions, 330, 335
in explosive power testing, 334
on gun projectiles, 290
nuclear blast wave effect, 366
vs. peak overpressure, 367
peak overpressure as, **367**
positive phase of, 346
subsonic, 331
suction phase of, 346
shock wave
on gun projectiles, 290
limitation to propeller driven aircraft, 276
to missile canisters, 290
short-range operation, 63
shortwave IR (SWIR), 101
sidelobe angle deception, 92
sidelobe cancellation and suppression, 96
sidelobes, 34
Sidewinder missiles, 138, **321**
sievert, 373
signal, 179
signal amplitudes and threshold levels, **36**
signal analysis capability, 84
signal integration time, 53
signal-processing techniques, 96–97
signal strength vs. target position in beam, **33**
signal-to-noise ratio (SNR)
and bandwidth selection, 36–37
and carrier frequency selection, 32
of conical scan systems, 73
defined, 36
and denial jammers, 87
figure of merit (FOM) and, 179
improvement of, 201–2
of monopulse trackers, 73
of multiple underwater sensors, 141, 204
of photovoltaic detectors, 141
simplified IR range equation, 145–46
simplified Laplace transform pairs, **217**
simplified radar range equation, 39–41
sine wave, **6**
sing-around principle, 161
single-lobe pattern, 33
skin drag, 244
skin echo, **71**
skin friction, 290
sky wave propagation, **23**
sky wave radars, 63–64
sky waves, 22–23
slant ranges
for gamma-ray doses, **381**, **382**
methods of determining, 74
for neutron doses, **382**
proximity fuzes used to determine, 318
for radiant exposures, **381**
for thermal radiation prediction, 380
SLBM (submarine launched ballistic missile), 309
small angle approximation, 137
smart warheads, 352
smokeless powders, 240
smoke warheads, 356–57
Snell's Law, 6, 11, **163**, **164**
sniperscope, 99
snorkel, 207
soft recoil systems, 266
soft recoil vs. conventional recoil, **266**
solar and lunar cycle illumination, **116**
solar illumination, 115
solid fuel ramjet, 282
solid-fuel rocket motor, **279**
solid geometry vs. plane geometry, 102
solid propellant, 278
solid-propellant configurations, **278**
soman, 385
somatic effects, 373–75
sonar equation parameters, 180
sonar equation terms, 179–86
absorption losses, 183
active source level, 180
detection threshold (DT), 179–80
directivity index, 185–86
noise level, 186
passive source level, 180
reverberation level, 186
scattering losses, 183–85
spreading loss, 181–83
target strength, 186
total propagation loss, 185
transmission loss, 180–81
sonic layer, 165
sonic layer depth (SLD), 165
sonobuoy deployment, **206**
sonobuoys, 205–6
sound channel propagation, **166–67**
sound intensity, 157
sound pressure levels, 156–57
sound sources and noise, 170–78
ambient noise, 175
biological noise, 176–77
cavitation, 172–73
flow noise, 171–72
hydrodynamic noise, 175
machinery noise, 171
ocean traffic, 176
quieting, 173
seismic noise, 176
self-noise, 170–71
submarine quieting, 174–75

sound sources and noise, (*Continued*)
surface quieting, 173–74
Wenz curve, 177
sound speed
and pressure, 159
vs. pressure, salinity, and temperature, **160**
and temperature, 159–60
sound speed equations, 160
sound speed profiles, 162–64
sound variation due to ship travel, **322**
sound velocimeter, 161
sound wave (ray), 163
source geometry, 108
source level (SL), 180
space waves, 23–24
spalling, 355
special cases, 292
special purpose projectiles, 291
special-purpose warheads, 355–57
antipersonnel warheads, 357
armor piercing warheads, 355–56
biological warheads, 356
chaff warheads, 357
chemical warheads, 356
cluster bomb units (CBUs), 357
fuel-air-explosive (FAE) warheads, 356
hard target penetrator warheads, 357
psychological warheads, 357
pyrotechnic warheads, 356–57
radiation warheads, 356
thermal warheads, 356
specific impulse of the propellant (solid fuel), 278
specific thrust of the propellant (liquid fuel), 278
spectral decomposition, 14
spectral distribution curves, **116**
spectral exitance of blackbody and graybody, **112**
spectral radiant exitance of blackbody, **111**
spectral response of eye (photopic) vs. night sky irradiation, **147**
spectral response of NVGs vs. night sky irradiation, **147**
spectral sensitivity of second- and third-generation image intensifier tubes, **150**
spectrum of pulsed Doppler radar, **50**
speed control, 213
speed control block diagram, **227**
speed of sound, 154, 160–62
speed of sound in water, 159–62
speed vs. self-noise levels, **172**
spherical array, **201**
spherical coordinate relationships with pitch and roll, **257**
spherical coordinate systems, **253–54**
spherical coordinate systems vs. rectangular coordinate systems, **255**
spherical spreading, **34**, 118, 167, **181**
spherical wave spreading, **10**
spin stabilization, 290
spot jamming, **89**
spreading, 15
spreading and divergence, 118–24
absorption, 119–21
atmospheric structure, 118–19
atmospheric transmission, 123
ocean propagation, 123
propagation in liquids and solids, 124
scattering, 121–22
turbulence, 122–23
spreading loss, 181–83
spring, **228**
stability, 333, 334
stabilization, 68
stabilized reference frame vs. unstabilized reference frame, **256**
stabilized weapon station frame, 249
stadimeter ranging, 137
staggered pulse repetition frequency, **49**
staggered pulse repetition frequency (PRF), **49**, 50, 58
stall, 284
stalled airflow, 284
standard deviation, 343, 344
Standard missile, 271
standard range, 180
standard speed of sound in water, 160
standard state, 339
stand-forward jamming, 88–89
standing sine wave, **7**
standoff distance, 354
stand-off jamming, 88
star check, 259
staring array, **139**
starlight, 116
starlight scope, 148
static thrust, 281
stationary targets, 46, 49
steady state error, 214, 231
steam-ejection-type system, 268
steered beam, 59
Stefan-Boltzman law, 111–12
step function, 219
steradian, 102–3, 311
Stinger missiles, 271
stochastic effects, 375
storage, 272–73
storage temperature, 333
stratosphere, 118
streamlining, 286
strength pressure curve, **263**
stridulatory sound, 177
string arrays, 138
subcaliber projectiles, 292
sub-clinical dose, 374
sub-critical mass, 363
submarine-launched naval moored mine, **296**
submarine quieting, 174–75
submarines
biological luminescence and, 115
ejector launchers in, 267–68
explosion near hull of, **360**
impulse propulsion, 276
laser detection of shallow, 124
mine laying by, 294–96
SLBM (submarine launched ballistic missile), 309
swirl behind, **207**
see also active sonar equation; figure of merit (FOM); sonar equation terms; underwater sound; underwater warfare systems
subsurface burst (underground), 370
subsurface communications, 209–10
suction phase of shock (blast) wave, 346
summing junction, 227, 228
sum pattern of two beams, **72**
sunlight, 115
supercritical chain reaction, 363
supercritical mass, 363
superheterodyne receivers, **85–86**
superheterodyning, 85
superimposition, 12
superRBOC (super rapid blooming offboard countermeasures), **93**
supersonic flight, 280
surface burst, 369–70
surface cutoff, 358
surface-duct propagation, **166**
surface quieting, 173–74
surface reverberation, 184
surface target vs. air target trajectories, **244**
sweep denial jamming, **89**
sweep jamming, 89
swirl behind submarine, **207**
switch horn, 324
switch-type fuze, **326**
symptoms of nerve agent poisoning, 385
synchronizers, 28, 198
synthetic aperture, **52**
synthetic aperture radar (SAR), 51–53
systematic error, 344–45
system block diagram, **217**, **227**, 234
see also block diagrams
system controllers or compensators, 233–35
composite system, 235–36
proportional control, 233
proportional-derivative (PD) control, 233
proportional-integral (PI) control, 233–34
proportional-integral-derivative (PID) control, 234–35
system performance, 233
system response, 219–20
system response and various damping ratios, **222**
system response examples, 223–27
system time responses
critically damped response, 220
effect of system components, 226–27
overdamped response, 20
undamped response, 222
underdamped response, 221–22
unstable non-oscillating response, 223
unstable oscillating response, 222–23
system transfer function, 218

T-62 tank, 131
T-62 tank signature, 131
tactical action officer (TAO), 1
tail control, 288
tailpipe radiance, 129–30
tamper, 364
tanks
M-60 tank signature, **131**
muzzle brake function on, 266

120 millimeter kinetic energy antitank projectile, **355**
shaped charge impact on, 355
T-62 tank, 131
T-62 tank signature, **131**
target acquisition gates, 77–78
target angle, 43
target detection, 3, 76–77, 144–45
target detection and recognition, 316–17
target detection device (TDD), 288, 317
target detection device (TDD) sensor, 318
target location, **257**
target location error, 312
target masking and modification (STEALTH), 94
target motion, 80
target motion analysis (TMA) techniques, 84
target position and return pulse amplitude, **70**
target position in beam vs. signal strength, **33**
targets and backgrounds, 125–30
target signal strength and beam axis, **33**
target size vs. frequency, 38
target strength (TS), 186, 188
target vulnerability, 342
temperature
for explosion test, **333**
of gun, 242
of powder, 241
and sound speed, 159–60
temporal coherency time, 134
tenth-thickness rule, 371
TERCOM maps, 310
terminal ballistics studies, 350
terminal engagement situation, **320**
terminal phase, 301
terrain contour mapping (TERCOM), 310
terrain contrast analysis, **117**
terrain/object reflectivity (albedos), **116**
terrestrial guidance, **310**
terrestrial guidance methods, 309–10
therapeutic dose, 374
thermal biological damage limits, 380
thermal detectors, 141–42
thermal jet engine, 277
thermal noise, 123
thermal pulse, 366, 368
thermal radiation, 109–14
absorptivity, 112
blackbodies, 111
emissivity, 112–13
Planck's law of blackbody radiation, 111
reflectivity, 112
Stefan-Boltzman law, 111–12
in tactical scenario, **110**
theory of, 109–10
thermal vs. selective radiators, 110–11
total emissive power per unit area, 111–12
transmissivity, 112
universal blackbody curve, 113–14
wavelength with peak radiation, 113
Wien's displacement law, 113
thermal radiation prediction, 380–81
thermal radiators, 110
thermal radiators output, **110**
thermal source, 123
thermal structure, 164–66
thermal vs. selective radiators, 110–11
thermal warheads, 356
thermite, 356
thermochemistry, 335
thermocline, 162
thermonuclear fusion, 365
3dB bandwidth, **47**
three-dimensional beam patterns, **195**
three-dimensional phased-array radar, **62**
three-dimensional radar scan pattern, **61**
three-dimensional reference frame, **249**
three-dimensional search radar using phased-array techniques in elevation, **60**
threshold levels, **36**
threshold SNR, 198
thrust, **277**
thrust-time curve, **278**
time and frequency domain representation, **9**
time-delay scanning, **56**
time domain, 217
time domain vs. frequency, 9
time fuzes, **317**, 324
time of flight (TOF), 237, 247
timer (synchronizer), 28
timer/command fuzes, 317
timer devices, 316
time relationship of blast wave vs. dynamic pressure, **346**
time relationship of blast wave vs. peak overpressure, **346**
tip vortex, 285
TNT
dynamic pressure for, **349**
heat of explosion for, 340
order of priorities for explosive product decomposition reactions, **336**
oxygen balance (OB%) of, 337–38
peak overpressure for, **349**
volume of products of explosion, 337
Tomahawk missiles, 281
tornadoes, 48
torpedo architecture, 292–94
torpedo force diagram, **292**
torpedo loading, 274
torpedo programmed path, **302**
torpedo tubes, 267
torpedo warheads, 360
total drag, 284, 286–87
total drag equation, 284
total emissive power per unit area, 111–12
total propagation loss, 185
towed array sonar by sections, **204**
towed array sonar system (TASS), 204–5
toxicity, 335
toxins, 389, **391**
track ambiguity, 78–79
track-breaking mode, **91**
track correlation and association, 78–79
tracker, 67
track file, 79
track file generation, 79
track file structure, **79**
track gate prediction, smoothing and positioning, 79–80
tracking, 2, 3
tracking algorithm, 76
tracking algorithm system equations, 80–81
tracking gates, 77–78
tracking line, 66, **67**
tracking systems, 66
track initiation, 79
track-on-jam techniques, 89
track via missile (TVM), 307
track-while-scan (TWS) radar, 75–82
acquisition and tracking gates, **77**
active tracking radar, 81–82
automatic detection and tracking (ADT), 76
integrated automatic detection and tracking (IADT), 82
target acquisition gates, 77–78
target detection, **76–77**
track correlation and association, **78–79**
track file generation, 79
track gate prediction, smoothing and positioning, 79–80
tracking algorithm, 76
tracking algorithm system equations, 80–81
tracking gates, 77–78
track initiation, 79
turn detection gates, 77–78
TWS system operation, **81**
trajectories, 301
tram bars, 258
transducer
crystal stack for, **193**
directionality of, **195**
transducer array, 196–97
transducer directivity, 194
transducer elements, **197**
transducer functions, **192**
transducer power, 195–96
transducers, 192–96
ceramic transducers, 193
crystal transducers, 192
direction determination, 195
hydrophones, 194
magnetostrictive transducers, 193–94
transducer directivity, 194
transducer power, 195–96
transfer equipment, 273
transfer functions, 229
transient radiation effects on electronics (TREE), 368, 377
transient response, 234
transition range, 182–83
translation, **257**
translational system block diagram, **218**
translational systems, **214**, 215–16
transmission
of atmospheric gases, **120**
through infinitesimal distance of atmosphere, **123**
transmission loss (TL), 156, 180–81, 182, 187
transmission loss anomaly, 185
transmission windows, 120–21, **123**
transmissivity, 112
transmit/receive (TR) modules, 62
transmitted power distribution, **27**

transmitter components, 37
transmitter power, 37
transmitters, 28, 87, 197
transmitting/receiving (TR) switch, 29
transponders, 69, 90
transverse magnification, **134**
transverse wave, 154
traveling waves
  amplitude, 7
  characteristics of, 5–9
  coherency, 6
  coherent energy, 6
  frequency, 5–6
  frequency vs. time domain, 9
  phase, 8–9
  velocity, 6
  wavelength, 6
TREE (transient radiation effects on electronics), 368, 377
trigger, 365
troposphere, 118
tropospheric ducting, 24
tropospheric scatter, 24
tropospheric scatter propagation, **24**
true bearings, 29, 43
turbofan engines, 280
turbojet engines, 280–81
turboprop engines, 280
turboshaft engines, 280
turbulence, 122–23
turbulent flow, **171**
turn detection gates, **77–78**
turning gate, **78**
two-element array
  end view of, **20**
  radiation pattern for, **21**
two-element frequency-scanned array, **57**
two person rule, 365
TWS algorithm flow diagram, **81**
TWS processing, **78**
TWS radar and crosseye (phase front distortion), 92
TWS system operation, 81

UH-1N Star Safire, 139
ultraviolet wavelengths, 100
undamped response, 222
undamped time response, **222**
underdamped response, 221–22
underdamped system root, **219**
underdamped time response, **221**
underground burst, **370**
underwater burst, **370**
underwater communications, 209–10
underwater electric potential (UEP), 323
underwater explosion, **358**
underwater mines, 321
underwater sound
  fundamental concepts, 154–59
  propagation paths, 164–69
  sound sources and noise, 170–78
  sound speed profiles, 162–64
  speed of sound in water, 159–62
underwater warfare systems, 192–212
  active systems, 196–200
  beamwidth, **194**
  directivity, 194
  nonacoustic detection, 207–9
  passive systems, 201–3
  transducers, 192–96, **193**, **195**
underwater warheads, 358–60
unitary chemical munitions, 388
units of measure of radiation, 373
unit step function, **219**
unit step input, 219
universal blackbody curve, 113–**14**, **144**
unstabilized reference frame vs. stabilized reference frame, **256**
unstabilized tracking subsystems, 68
unstabilized weapon station frame, 249
unstable non-oscillating response, **223**
unstable oscillating response, 222–**23**
unwanted signal rejection (narrow bandpass), 84
up Doppler, 200
upper frequency limit, 205
upslope sound propagation, 169
urticants, **387**
U.S. Naval Research Laboratory (NRL), **71**

vacuum trajectories vs. atmospheric trajectories, **245–46**
variable depth sonar (VDS), 203
variable geometry inlets and exhausts, 280
variable homing logic, 303–4
variable paths, 301
variable time (VT) fuzes, 317
vector sum diagram, **302**
velocity, 6
velocity discrimination, 46
velocity feedback, 74
velocity in the line-of-sight (LOS), 42–43
velocity limitations of combination radar systems, 49–50
velocity of jet tip, 354
vertical beam of cylindrical transducer array (declination angle), **197**
vertical launch system (VLS), **271**
vertical polarization, **17**, 18, 22
very high frequency (VHF), 23
very low frequency (VLF), 23
vesicants, 36, 386
visible electro-optical (EO) detection, 109
visible light, 100–101
visible radiation, **101**
visible wavelengths, 100
visual, 207
visual photometry, 102
VLAD (vertical thin-line array DIFAR), 206
volatility, 334
volume elasticity, 159
volume of products of explosion, 337
volume reverberation, 184
VX, 385

wake, 290
warhead characteristics, 345
warhead initiation, 317
warhead section, 294
warhead size, 378
warhead to target proximity, 341
waterfall display, **203**
wave drag, 245
wave equation, 162
waveforms with delay in phase, **57**
wavefront, 17, **54**
wave guide, **29–30**
wave guide field diagram, **30**
wavelength
  defined, 6
  of maximum radiation, 366
  with peak radiation, 113
  vs. target size, 38
  of visible and infrared radiation, **101**
wave motion, spring and mass model, **154**
wave polarization, 17
wave propagation, 9–15
  basic phenomena, 9–10
  diffraction, 11–12
  Fourier analysis principle, 14–15
  interference, 12–14
  refraction, 10
weapon architecture, 282
weapon launch lead angle, 275
weapon lethality, 342
weapon propulsion, 276
weapons of mass destruction (WMDs), 361
weapons selection, 2, 3
weapons systems, 3
weapon station reference frames, 249–50
weapon system alignment, 258–61
  aircraft boresighting, 259–61
  alignment, 258–59
  collimation, 259
weapon trajectory, **243**
weather, 36
weather radars, 37, 47–48
Wenz acoustic ambient noise in the ocean curves, **177**
Wenz ambient noise curves, **178**
Wenz curve, 177
Wenz ocean traffic categories, **176**
white-hot polarity, 123, **127**
wide area augmentation system (WAAS), 312
wide dynamic range, 84
wide spectrum surveillance, 84
Wien's displacement law, 113, 366
wind, 246
wing aspect ratio, 285
wing control, 288
wingtip vortices, **285**
wire-guided command system, 305
wrap-up ratio, 58

X-ray examination dose, 373

yield, 364

zero-length launchers, **270**
zero oxygen balance, 337

About the Editor

Lt. Cdr. Craig Payne, USN, is a career surface warfare officer who has served as a tactical action officer on frigates, destroyers, and an aircraft carrier. He has also served as the battle watch captain of a NATO Destroyer Squadron staff. Lieutenant Cammander Payne was commissioned through the NROTC program at the University of Washington where he earned a bachelor's of science in Chemical Oceanography. After his first sea tour, he earned a master's of science in Applies Science (ASW) from the Naval Postgraduate School. He and his very supportive wife Cindy have three children: James, Stephanie, and Cameron. Lieutenant Commander Payne is currently a master instructor and coordinator of the Naval Weapons Systems courses at the U.S. Naval Academy.

The Naval Institute Press is the book publishing arm of the U.S. Naval Institute, a private, nonprofit membership society for sea service professionals and others who share an interest in naval and maritime affairs. Established in 1873 at the U.S. Naval Academy in Annapolis, Maryland, where its offices remain today, the Naval Institute has members worldwide.

Members of the Naval Institute support the education programs of the society and receive the influential monthly magazine *Proceedings* and discounts on fine nautical prints and ship and aircraft photos. They also have access to the transcripts of the Institute's Oral History Program and receive discounted admission to any of the Institute-sponsored seminars offered around the country. Discounts are also available to the colorful bimonthly magazine *Naval History*.

The Naval Institute's book publishing program, begun in 1898 with basic guides to naval practices, has broadened its scope to include books of more general interest. Now the Naval Institute Press publishes about seventy titles each year, ranging from how-to books on boating and navigation to battle histories, biographies, ship and aircraft guides, and novels. Institute members receive significant discounts on the Press's more than eight hundred books in print.

Full-time students are eligible for special half-price membership rates. Life memberships are also available.

For a free catalog describing Naval Institute Press books currently available, and for further information about joining the U.S. Naval Institute, please write to:

Member Services
U.S. Naval Institute
291 Wood Road
Annapolis, MD 21402-5034
Telephone: (800) 233-8764
Fax: (410) 571-1703
Web address: www.navalinstitute.org