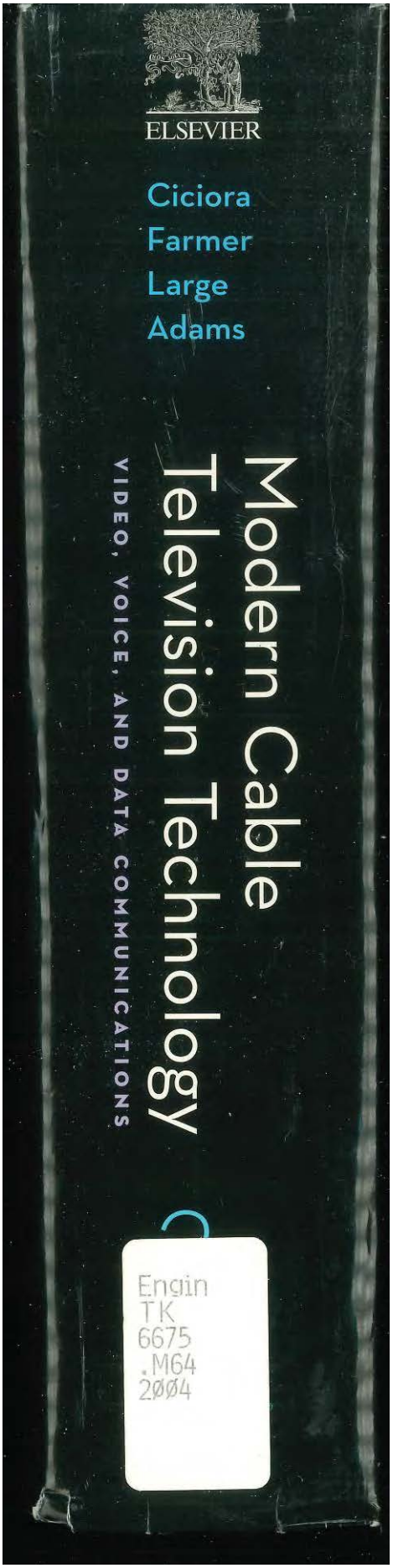
Walter Ciciora
James Farmer
David Large
Michael Adams

Modern Cable Television Technology

VIDEO, VOICE, AND DATA COMMUNICATIONS

2nd Edition

Morgan Kaufmann


ELSEVIER
Ciciora
Farmer
Large
Adams
Modern Cable Television Technology
VIDEO, VOICE, AND DATA COMMUNICATIONS
Engin
TK
6675
.M64
2004

Modern Cable Television Technology

VIDEO, VOICE, AND DATA COMMUNICATIONS

Walter Ciciora, *Consultaant*
James Farmer, *Wave7 Optics, Inc.*
David Large, *Altrio Communications*
Michael Adams, *Broadband Semantics*

Second Edition

For engineers and technicians engaged in the modern cable telecommunication environment, the second edition is the essential guide to the latest developments and preparation for what is to come.

—From the Foreword by Archer Taylor, retired, Co-founder, Malarkey-Taylor Associates

Wow! What an outstanding contribution to the entire telecommunications industry this book has become! It's an absolute "must have" for any serious engineering professional in this or related industries.

—Chris Bowick, Senior VP of Engineering and Chief Technical Officer, Cox Communications

This 'post-Modern' 2nd edition of their classic text is the new reference standard for the Cable industry.

—Dan Pike, CTO, GCI Cable

I have certain books in my library that are special, they are so concisely written that they always answer a question or promote clear thinking. These books, from another era, clearly explain radio physics and television; they are now joined by *Modern Cable Television Technology* as it too is a classic work on the art.

—Ted Hartson, Principal, Scottsdale Television Labs

Fully updated, revised, and expanded, this second edition of ***Modern Cable Television Technology*** addresses the significant changes undergone by cable since 1999—including, most notably, its continued transformation from a system for delivery of television to a scalable-bandwidth platform for a broad range of communication services. It provides in-depth coverage of high speed data transmission, home networking, IP-based voice, optical dense wavelength division multiplexing, new video compression techniques, integrated voice/video/data transport, and much more.

Intended as a day-to-day reference for cable engineers, this book illuminates all the technologies involved in building and maintaining a cable system. But it's also a great study guide for candidates for SCTE certification, and its careful explanations will benefit any technician whose work involves connecting to a cable system or building products that consume cable services.

FEATURES

- The much-awaited second edition of an award-winning book, written by leading figures in the cable industry.
- Organized to "follow the plant" from signal creation, through multiplexing, transmission, and, finally, reception and processing within consumer's premises.
- Focuses on the practical, not the theoretical, and explains concepts and techniques using a minimum of mathematics.
- Covers both analog and digital signals, as well as coaxial and fiber-optic broadband distribution systems.
- Discusses system architecture in detail, including considerations relating to digital fiber modulation and network reliability.
- Explores a wide range of customer interface issues, including analog and digital video reception, consumer electronics, and home networks.

ELSEVIER
MK
MORGAN KAUFMANN PUBLISHERS
AN IMPRINT OF ELSEVIER
www.mkp.com

ISBN: 1-55860-828-1
90000
9 781558 608283

Networking/
Communications

The Morgan Kaufmann Series in Networking

Series Editor, David Clark, M.I.T.

Modern Cable Television Technology: Video, Voice, and Data Communications, 2e
Walter Ciciora, James Farmer, David Large, and Michael Adams

Bluetooth Application Programming with the Java APIs
C Bala Kumar, Paul J. Kline, and Timothy J. Thompson

Policy-Based Network Management: Solutions for the Next Generation
John Strassner

Computer Networks: A Systems Approach, 3e
Larry L. Peterson and Bruce S. Davie

Network Architecture, Analysis, and Design, 2e
James D. McCabe

MPLS Network Management: MIBs, Tools, and Techniques
Thomas D. Nadeau

Developing IP-Based Services: Solutions for Service Providers and Vendors
Monique Morrow and Kateel Vijayananda

Telecommunications Law in the Internet Age
Sharon K. Black

Optical Networks: A Practical Perspective, 2e
Rajiv Ramaswami and Kumar N. Sivarajan

Internet QoS: Architectures and Mechanisms
Zheng Wang

TCP/IP Sockets in Java: Practical Guide for Programmers
Michael J. Donahoo and Kenneth L. Calvert

TCP/IP Sockets in C: Practical Guide for Programmers
Kenneth L. Calvert and Michael J. Donahoo

Multicast Communication: Protocols, Programming, and Applications
Ralph Wittmann and Martina Zitterbart

MPLS: Technology and Applications
Bruce Davie and Yakov Rekhter

High-Performance Communication Networks, 2e
Jean Walrand and Pravin Varaiya

Internetworking Multimedia
Jon Crowcroft, Mark Handley, and Ian Wakeman

Understanding Networked Applications: A First Course
David G. Messerschmitt

Integrated Management of Networked Systems: Concepts, Architectures, and their Operational Application
Heinz-Gerd Hegering, Sebastian Abeck, and Bernhard Neumair

Virtual Private Networks: Making the Right Connection
Dennis Fowler

Networked Applications: A Guide to the New Computing Infrastructure
David G. Messerschmitt

Wide Area Network Design: Concepts and Tools for Optimization
Robert S. Cahn

For further information on these books and for a list of forthcoming titles, please visit our website at http://www.mkp.com

Senior Acquisitions Editor Rick Adams
Associate Editor Karyn Johnson
Senior Project Manager Angela Dooley
Cover Design Ross Carron Design
Cover Photographs Fred Spencer and Carolyn Terry
Composition and Illustrations Integra Software Services Pvt. Ltd, Pondicherry, India
Copyeditor Elliot Simon
Proofreader Fran Pulver
Indexer Northwind Editorial
Interior printer The Maple-Vail Book Manufacturing Group
Cover printer Phoenix Color

Morgan Kaufmann Publishers in an Imprint of Elsevier
500 Sansome Street, Suite 400, San Francisco, CA 94111

This book is printed on acid-free paper.

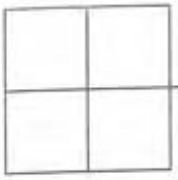
Library of Congress Catalog Number: 2003107478
ISBN: 1-55860-828-1

For all information on all Morgan Kaufmann publications
visit our website at *www.mkp.com*

Printed in the United States of America

03 04 05 06 07 5 4 3 2 1

Contents



Chapter 1
Introduction to Cable Television

1.1 Introduction

Cable television is an industry and a technology that has outgrown its historical name. Modern "cable television" networks are used to provide a wide range of services, including analog and digital video, digital audio, high-speed data, and telephony.

The essential distinguishing characteristics of cable television networks are that they include broadband (typically 0.5–1 GHz of total bandwidth), highly linear distribution systems designed to carry many modulated radio frequency (RF) signals with a minimal amount of mutual interference between a central point and many customers, where signals are delivered via coaxial cables to and from terminal equipment. Because of these characteristics, the networks are service-agnostic to the extent that they will carry any information that can be modulated on a compatible RF carrier. Modern cable television networks are almost always two-way, use optical fiber extensively, and are segmentable so as to allow simultaneous frequency reuse in various network sections.

Historically, the cable television business was based exclusively on delivery of television programming, and it has been very successful in that regard. As of 1999, nearly 97% of U.S. television households had cable television service available, and approximately 66 million households subscribed to at least the lowest tier of video service, representing almost 67% of U.S. television households.[1] Those levels have changed only slowly over the past few years.

Because cable television has been so successful and has enjoyed such vigorous growth and acceptance, it has spawned video competitors, including prerecorded media, direct broadcast satellite (DBS), video streaming over the Internet, as well as the interest of the telephone industry. Increasing revenue streams from connected households, made possible by multiple service offerings, has also changed the economics of the industry sufficiently that customers in some markets now have a choice between two cable television operators who have constructed parallel distribution networks serving the same homes. In the common terminology of the cable industry, the company building the second network is known as an *overbuilder*. Overbuilders may offer service as a second, franchised cable television operator or as an *open video system* (OVS) operator.

Taking advantage of new technologies and, in particular, the falling prices of electro-optical components, these networks typically use optical fibers to carry signals closer to subscribers than legacy cable operators, in some cases all the way to subscriber homes.

In the world of high-speed data communications, and Internet access in particular, services offered by cable operators have codeveloped with other wired and wireless options. For residential users, Internet access was historically provided almost exclusively through dial-up modems directly to *Internet service providers* (ISPs). Many applications, however, run discouragingly slowly at the data rates that are possible through standard telephone connections, and this has led cable companies to develop connection services that are 10–100 times faster. In the competition among broadband service providers, cable has outsold its competitors by about 2:1, with a telephone technology known as *digital subscriber line* (DSL) providing the strongest competition to date. Competing satellite and wireless terrestrial data transport technologies are still developing market share.

In offering voice telephone service, it is the cable operator who is the overbuilder, since residential telephone service was available to virtually every household in the United States before cable television operators entered that market segment. When offered by cable television operators, telephone service is regulated by the same agencies who regulate the incumbent telephone companies, leading to completely different regulatory authorities for services that share the same physical network.

Cable-offered telephony comes in at least two technical versions and two product classifications. Initially, telephone offerings utilized dedicated signal-processing equipment that was carefully engineered to meet the high reliability expectations of this service. As of early 2003, that type of equipment was still used to service the large majority of installed telephone customers. The newest version, known as *voice-over-IP* (VoIP), shares terminal equipment with the high-speed data service, offering the potential for reduced equipment costs.

Regardless of how the signals are handled technically, cable operators may offer a *primary-line* service or only a *secondary-line* service. Primary-line service competes directly with the incumbent telephone operator for all residential telephone business, but it requires that the cable network and equipment meet the reliability standards to be the "life-line" communications link for customers in the case of an emergency (while not mandated by regulatory agencies, this voluntary requirement is generally defined as a service that is available at least 99.99% of the time). Secondary-line service cedes the first line in each home to the incumbent but competes for additional lines. The assumption is that the availability needn't be as high, saving the cable operator capital upgrade cost and allowing it to offer a lower price.

In the following chapters, you will gain a solid understanding of the technologies required to deliver broadband services to and from homes. You will see how the pieces fit together to make up a complete system for the transmission of

information and entertainment choices to consumers. If yours is a related business, you will better understand how it fits with the cable industry. If you are already knowledgeable in some aspects of the broadband networks used by cable, this book will fill in the gaps.

1.2 Technology-Related Regulatory Issues

Cable's high visibility has attracted the attention of regulators and legislators. Important public policy issues are involved. Understanding what cable television is, how it works, and its economics will help decision makers in these areas. This understanding will also aid network designers in determining which technologies are appropriate for cable television applications and which are not. One aim of this book is to aid in that understanding.

The issues are fundamentally different from those that govern incumbent telephone companies, broadcasters, or DBS operators because of the nature of the cable television business model. Unlike telephone companies, which are fundamentally providers of bandwidth but have no interest in content, and DBS operators, which are fundamentally purveyors of content and for which the network is just a private delivery mechanism, cable companies are involved in construction, network operations, bandwidth leasing, passive retransmission of certain signals, resale of certain programs and signals, equipment leasing, and program generation. These sub-businesses are the result of cable's need to define itself over the years in a very entrepreneurial way and make it hard to pigeonhole it into a defined niche. Following are a few of the regulatory issues that are affected by the technology of cable systems.

1.2.1 Access to Networks by Unaffiliated Third Parties

Current regulatory policy allows unaffiliated competitive local exchange carriers (CLECs) to offer telephone and data services over physical plant constructed by incumbent telephone companies and to access incumbent telephone company facilities for the purpose of installing the necessary equipment. In some cases, signals from the two companies may share the same physical wires between a telephone switching office and customers' homes. A recurring regulatory question is whether, and to what extent, access to cable systems should be granted to such third parties. Should access be granted at the physical level (shared wires), at the bandwidth level [allocated radio frequency (RF) channels in the network], or at the service level (where the third party's signals are processed by equipment owned and operated by the incumbent cable operator)?

The question of third-party access potentially arises in connection with all types of cable television services, often as a result of the vertical organization of cable companies, which not only build and operate networks but may also invest in companies providing video programming, operate their own ISPs, resell

long-distance telephone service, lease various types of terminal equipment, and have other relationships that may cause them to favor some suppliers over others.

The technical issues are much more complex in the cable television world than in the telephone world because cable systems are not passive, but, rather, include electro-optical and RF active equipment whose operation requires very tight control over signal parameters. Furthermore, cable system operators are responsible for inadvertent radiation of their signals, which can be caused by poor workmanship on the part of a third-party installer.

1.2.2 Cable Operators' Responsibilities in Transition to Digital Broadcast Television

The eventual transition of over-air broadcast television from analog to digital (including both standard and high-definition formats) is a certainty, though the timing is not. The practical issues are many.

- Broadcasters are reluctant to produce high-definition television (HDTV) in quantity because of the high costs involved for a currently-very-limited viewing audience, yet HDTV programming in sufficient quantity will be required to encourage consumers to purchase digital receivers.
- The costs of digital receivers remain high due to limited demand as well as inherently higher production costs.
- DBS and cable television operators offer both standard-definition television (SDTV) and a limited quantity of HDTV programming, but using formats that are incompatible with each other and with over-air broadcast digital. Even among cable network operators and between satellite operators, receivers are not currently interchangeable. Furthermore, HDTV programs use four to six times the bandwidth of SDTV programs, so any further migration to HDTV on the part of DBS and cable operators will inevitably be at the expense of reduced program variety, given a finite amount of available bandwidth.
- Given that most households in the United States receive their television programming from a cable television system, broadcasters maintain that a successful transition is dependent on cable operators' carrying their analog and all their digital signals during the transition period. Given their limited bandwidth, the relative scarcity of deployed HDTV receivers, and the duplication of much of that programming, cable operators believe that this is an unreasonable burden on their limited bandwidth and an imposition on the editorial control of their systems; they point out that the immediate result will be to reduce programming choice for their customers and to deny access to some nonbroadcast programmers.
- For the consumer, the result has been that those wishing access to digital programming typically purchase a display-only receiver, then mate it with separate satellite, cable, over-air, and/or physical media players, as required.

1.2.3 Market Power Issues

Although not a legal monopoly, cable television systems, in total, serve about two-thirds of U.S. households with at least television service. This has led some regulators to regard them as *de facto* monopoly providers and to enact regulations to limit their power to control various aspects of the marketplace.

Consumer Terminal Equipment

The Cable Act of 1996 led to Federal Communications Commission (FCC) regulations in 1998 that required that any required terminal equipment be accepted from unaffiliated third parties by July 2000. Note that this regulation applies only to cable television (video) services, not to telephone services, which are provided under separate regulatory framework. The status of data services is still unclear.

The technical issues in successfully moving terminal equipment into the retail market are many and include the following.

- The requirement that such devices operate under the secure control of the network operator in allowing selective access to programming present at its input terminals. In practical terms this requires that the decryption/descrambling system, or at least a portion of it, be supplied by and controlled by the network operator.
- In order to meet policy needs to support the transition of the over-air broadcast industry to digital, as well as to provide reception of both broadcast and narrowcast programming over the cable network, the terminals must be capable of receiving, demodulating, and decrypting multiple types of signals.
- In order to protect the cable network, any upstream signals (transmitted from the customer back to the operator for control of interactive services) generated by the receiver must be tightly controlled and conform to industry standards for interactive communications. As of mid-2003, no such complete set of standards existed.
- In order to protect the cable network from destructive interference and to protect over-air communications services from interference, such receivers must be adequately shielded to prevent ingress or egress of signals.
- In order to ensure operability in any cable system, the reception performance must be consistent with the signals delivered by advanced cable system architectures.
- In order to ensure adequate functionality, the receivers must include an adequate set of features to allow the required subscriber interactivity, such as purchasing on-demand programming or executing e-commerce transactions.

- Even if all these goals are met, the result will be that cable operators will be inhibited from offering new and innovative services if those services require an upgrade or enhancement to subscribers' terminal devices.

Although multiyear intensive negotiations between the cable television and consumer electronics industries produced the outlines of specifications for "digital-ready" television sets that could be sold through retail outlets and that would function as both an over-air and cable receiver,[2] the reluctance of the FCC to set and enforce adequate labeling, feature, and performance standards left the effort in limbo.

A more recent private agreement between seven major cable operators and 14 consumer electronics manufacturers covered only noninteractive receivers and, as of the writing of this book, had not yet been adopted by the FCC.[3]

The cable industry is rapidly moving all premium video services from older analog scrambling technology to all-digital transmission and encryption. Therefore, it is generally considered prudent to concentrate on the regulation of digital video and to let proprietary analog scrambling systems die a natural death.

Greater success has been reached in high-speed data service, where a durable industry-wide standard has been adopted for terminal equipment. The Data Over Cable Systems Interface Specification (DOCSIS) was originally a proprietary document, but it has since been recognized by the International Standards Organization (ISO). This evolving effort is in its third major release, and devices meeting DOCSIS are usable in the vast majority of cable systems in the Americas. A major contributor to the success of this effort has been the activities of the U.S. research and development consortium, Cable Television Laboratories (CableLabs), acting as the official certifying agency for new equipment. Perhaps the strongest indicator of the success of DOCSIS is that the price of data modems has dropped by a factor of four in about two years, while the price of proprietary cable television video terminal equipment has declined by less than 50%.

Access to Programming by Competing Network Operators

At the other end of the network, regulators are concerned that cable's market power may allow operators to effectively control access to video content that is, after all, the core product customers purchase. This control can take the form of exclusive deals covering certain geographic areas or unreasonably unequal rates. In many cases, large multisystem operators (MSOs) are vertically integrated to the extent that they hold major investments in programming suppliers and thus influence how and to whom the programs are made available.

In the interest of encouraging competition, current regulations forbid exclusive access to any programming that is distributed via satellite, allowing both DBS operators and overbuilders access to the vast majority of programming.

Service Bundling and Buy-Through

Cable operators often bundle certain services (that is, offer certain services only in combination with other services) or require customers to purchase certain services in order to have access to purchasing additional services. In some cases, this is required by regulations (such as the requirement that the lowest level of video service include all the local broadcast stations) and in some cases it reflects available technology (such as the inability to reasonably block the reception of some services while allowing access to others). In other cases, however, it reflects marketing decisions, contracts with certain program suppliers or strategic alliances that may not result in choices customers would have preferred. Regulators, therefore, need to know what is technically reasonable, as well as to assess what is a reasonable degree of constraint on cable operators' rights to control their product.

One example that reflects the development of the business is the bundling of ISP service with high-speed data transport service. In some cases, a high-speed ISP company actually operates the Internet access business for the cable operator, and the combined revenue for modem rental, ISP services, and transport are what supports the business. Allowing subscribers to choose a different ISP may be technically difficult, depending on the way the system is configured, but it may also cause a redefinition of a formerly integrated business segment.

1.2.4 Other Technical Regulatory Issues

Numerous other regulatory issues are related to technology, including the following.

Delivered Signal Quality

Current regulations cover only analog video signals and do not address any digital services or two-way communications. The analog regulations are intended to ensure a reasonable subjective quality of television viewing.

System Shielding Integrity

Cable systems utilize a wide range of RF frequencies, including those used by other radio systems for open-air transmission. To ensure a lack of interference to those licensed users of the spectrum, the FCC has enacted and enforces strict signal leakage standards.

Must-Carry/Retransmission Consent Issues

In addition to the digital transition issues discussed earlier are a host of other issues related to cable carriage of over-air broadcast signals, including which

signals must be carried, where they are placed in the RF spectrum, what quality is required, whether format conversions are allowed or required, and what is fair compensation for use of the programming. A related issue is cable's obligation for carriage of over-air broadcast signals relative to that of DBS operators.

Public, Educational, and Governmental (PEG) Access Issues

Franchised cable television operators currently negotiate with franchisors (which may be local cities, states, or a combination thereof) regarding access by various public entities to channels on the cable system. The usual use of such channels is for distribution of local programming (such as transmission of city council meetings to local residents), but sometimes facilities are demanded for internal communications needs, such as two-way data communications between civic buildings.

Access by Cable Operators to the Utility Poles and the Public Rights-of-Way

A cable franchise or OVS agreement provides cable operators with access to the public rights-of-way. Overhead facilities within those rights-of-way, however, are usually controlled by third parties, such as incumbent telephone operators, who may not wish to encourage construction of networks that could compete with their core businesses. One interest of regulators, therefore, is in ensuring that access to existing pole lines is granted on a compensatory, but reasonable, basis to cable operators.

A related issue is access to ground-level and below-ground rights-of-way. Regulators are concerned about the appearance of neighborhoods and the integrity of their streets, while operators need prompt access under financially reasonable conditions to build networks. It is essential that regulators understand what is technically feasible in addressing these issues.

1.3 The Development of the Cable Television Industry and Its Services

Cable television systems were not originally intended to be general-purpose communications mechanisms. Only recently has cable's role expanded to include nonanalog video services. Even then, these services are designed as much as possible around the performance required for analog video transport—for instance, they get the benefit of the high carrier-to-noise ratios required for analog video transport but are also generally constrained to operate within the 6-MHz channelization scheme that is a byproduct of the over-air frequency assignments.

The earliest cable systems were designed solely to be able to transport a common spectrum of analog television signals from a central point, called the

headend, to each subscriber's home, where they would connect directly with the television receivers (in the parlance of cable operators, this is referred to as the *downstream* direction). The systems and receivers were designed for signals conforming to the U.S. television standard, called *NTSC*, after the organization that created it in 1941, the National Television Systems Committee. This black-and-white television standard was modified in 1953 to provide compatible color information to color receivers and again in 1984 to add compatible stereo sound.

The original business case was based on the delivery of broadcast television to areas where they were not receivable with reasonable off-air antennas. In rural areas this was usually due to the lack of available signals, whereas in cities multiple reflections from tall buildings created many reflections ("ghosts"). These early systems were called *community antenna television* (CATV) and consisted of sometimes-elaborate off-air antennas constructed where required for good reception, coupled to broadband distribution systems that carried the signals to customers' homes.[4] Although other transmission lines were used in some early systems, coaxial cable soon dominated. Because of the relatively high loss of then-available cable, the technical progress of the early industry was paced by improvements in the technical properties and reliability of repeater amplifiers, cable loss, and such accessories as subscriber taps and connectors.

By the late 1960s, nearly all the areas of the United States that could benefit from a CATV system had been served. In the mid-1970s, the launch of satellite delivery of signals to cable systems changed the business model from simply extending the reach of over-air terrestrial broadcasters to dramatically increasing program choices for customers and actively competing with the broadcasters for audience share. The seminal event was the cablecast by the pioneering movie channel Home Box Office (HBO) of the October 1975 Ali–Frazier heavyweight boxing match from Manila.

Three new categories of cable-exclusive channels came into existence: (1) "superstations," local over-air broadcasters using satellite transmission to gain a nationwide audience; (2) specialized channels for news, sports, weather, education, shopping, and so on; and (3) movie channels, such as HBO, which sparked new excitement in the business. With this new programming, cable was able to tap into many new revenue streams and to offer television programming tailored to customers' wishes. Within a few years, customer video offerings typically included the following.

- A basic level of service typically including local broadcast stations, franchise-mandated local access channels, and perhaps a few additional channels
- One or more enhanced tiers of satellite-delivered specialized channels for additional monthly fees
- A choice of commercial-free movie and special-event channels, for an additional monthly fee per channel

- The ability to purchase access to individual showings of movies or special events, known as pay-per-view, or PPV

New, nonsubscriber, revenue streams came from sales of commercials inserted in satellite-delivered specialized channels and from a share of the sales revenues earned from "home shopping" channels that were usually carried as part of the basic service.

The availability of many possible television channels, and the possible revenue to be gained by carrying them, led to a continuous expansion in the bandwidth of cable systems. The earliest systems were limited to all or part of the 12 VHF broadcast channels (2–13). The first expansion utilized the spectrum between the top of the FM band at 108 MHz and the bottom of channel 7 at 174 MHz. After that, channels were added above channel 13. Over the years the maximum frequency has increased in steps from 216, through 270, 300, 330, 400, 450, and 550, to 750 and sometimes 870 MHz. At least two experimental systems were built with a maximum frequency of 1,000 MHz. Appendix A lists the now-standardized channel numbers and frequency ranges. The need to periodically upgrade distribution systems is one factor that makes cable television a capital-intensive business.

Delivering multiple levels of programming also required the development of various methodologies for selectively controlling access to individual channels, groups of channels, and individual programs. The need to provide and periodically upgrade these terminal devices is a second factor that drives cable's capital requirements.

Since the early 1990s, cable operators have actively tried to capitalize on their network investment to deliver additional services, both video and nonvideo. While there are exceptions, the general characteristics of these new services are that they:

- Require two-way communication between customers and the headend
- Are characterized by communications between the headend and individual customers
- Utilize digital transmission

One of the first additional services was the delivery of larger quantities of video that would fit in the existing or anticipated bandwidth of cable networks if transmitted in standard NTSC analog format. Digitally compressed video with quality comparable to analog video was developed at approximately the same time (and by overlapping groups of people) as high-definition television in the United States.[5] It has since become the enabling technology for DBS, as well as the emerging technology of the U.S. over-air broadcast stations. What it did for cable television is to allow six to twelve digital programs to fit in the same bandwidth as a single analog program, coupled with improved encryption

security. This, in turn, enabled a new level of service, known as *near video on demand* (NVOD), whereby the most popular movies were offered with frequent start times (typically 15-minute to half-hour intervals).

Converting cable networks to handle two-way communications allowed the best use of this increased program capacity by offering an easy means to purchase individual events using upgraded set-top boxes that included upstream data transmitters. Although coaxial amplifiers had been designed to allow conversion to two-way communications since the 1970s, the upstream modules were seldom installed, lacking an incremental revenue stream to support the hardware investment.

Conversion to two-way in large systems was, however, problematic because of the accumulated noise buildup in the upstream direction in large all-coaxial networks with cascades of 20 or more series-connected amplifiers. Beginning in about 1990, however, the use of linear fiber-optic technology allowed much shorter cascades and broke large systems into many smaller systems, independently fed by fiber. This not only made two-way communications practical, but had the incidental advantage that the programming fed to each coaxial subnetwork could be customized to the customers fed from that section of plant.

This combination of developments, driven initially by revenue opportunities in delivering more video choice, led to the offering of additional services that depended on reliable two-way communications and network segmentation. Meeting the requirements for these services has lead to new ways of assessing systems. In particular, each of these services depends on *per-user bandwidth* rather than *total system bandwidth* as a primary measure of system capabilities.

- High-speed data and Internet access in particular have become a fast-growing business segment for the cable industry. Cable companies provide two-way data paths between each user and the Internet, but they also may lease modems and sell retail services, such as Web hosting and e-mail. As noted earlier, the industry has been very successful in this market, outselling their competitors by a significant margin.
- True video-on-demand (VOD) differs from NVOD in that users can select movies from a library, control start times, and, during playback, have VCR-like control — pause, rewind, fast forward, and stop. It also differs from any of the previously listed video services in that a user-selected video event is sent privately to a single customer and thus requires dedicated bandwidth for the duration of the playback. At the end of 2002, VOD had moved beyond trials and wide-scale deployments were just beginning. One attraction to VOD is that it can't be duplicated by DBS, which can neither segment its audience sufficiently, nor support the required two-way communications.
- The move to offer voice telephone services presented the greatest technical challenge to cable operators. Although the bandwidth requirements are modest, meeting the BellCore reliability standard of 99.99% service availability was beyond the capabilities of most cable networks. Those

operators offering a primary-line service generally invested additional capital in "hardening" their networks, primarily through installation of more redundant field power supplies, but also committed to higher staffing levels.

Clearly, this is not an exhaustive list. Any service that can be modulated onto an RF carrier of reasonable bandwidth can potentially share a cable television network. In the future, we can expect to see e-commerce applications, Internet access from television receivers, interactive gaming, and a host of other applications.

In the remainder of this chapter we will discuss, at a high level, the evolving networks that supported this business model.

1.4 Cable Network Design

Since cable television originally was not a general-purpose communications mechanism but, rather, a specialized system for transmitting numerous analog television channels in a sealed system, the topology, or layout, of the network was customized for maximum cost efficiency in performing that function. The architecture that evolved before the introduction of fiber optics was called *tree-and-branch*.[6]

There are five major parts to a classical cable system: (1) the headend, (2) the trunk cable, (3) the distribution (or feeder) cable in the neighborhood, (4) the drop cable to the home and in-house wiring, and (5) the terminal equipment (originally as simple as the subscriber's television set but sometimes with an intervening converter in case the tuning range of the receiver was not sufficient).

The headend is the origination point for signals in the cable system. Signals are received there from off-air antennas, parabolic satellite antennas, terrestrial microwave links, or coaxial links from local sources. Sometimes programming is generated locally. Each channel is locally processed and converted, or modulated, to the proper television channel for distribution to customers. Then the channels are combined into a complete spectrum through a process known as *frequency-division-multiplexing*, or FDM, before insertion into the coaxial distribution system.

The distribution system has the task of delivering the complete FDM spectrum to each customer's home while meeting certain quality standards. At various points in cable's history these standards have been voluntary, have been set by various franchising agencies, or have been preempted and set by the Federal Communications Commission. They are also driven by the need to compete with various other television entertainment sources, including pre-recorded media. Currently, FCC rules govern the minimum acceptable performance, and are summarized in Chapter 15 (Table 15.1).

It is worth noting at this point that, unlike some other communications systems, cable operators generally measure wideband thermal noise levels separately from levels of discretely identifiable signals and distortion products. They do so because such noise sources arise from different physical mechanisms and affect picture quality in different ways. Thermal noise appears in analog

television pictures as "snow," while discrete signal interference appears as parallel patterns of lines with various spacings and angles. Most commonly, thermal noise is a function of the number and input levels of repeating amplifiers, whereas discrete interference can arise from intermodulation products or over-air signals that "leak" into the cable system or subscriber's television receivers.

Common design parameters for distribution systems (not including any degradation that may occur in terminal equipment) call for a carrier-to-thermal-noise ratio (C/N) of 48–49 decibels (dB) and carrier-to-discrete-signal-interference ratio of 53 dB. These target values are consistent with subjective data compiled in 1991 in a project sponsored by CableLabs. That work, supervised by Dr. Bronwen Jones, indicates "perceptible, but not annoying" thermal noise at a C/N of about 47–50 dB. The "slightly annoying" level was within about 1 dB of 41 dB.[7]

The most efficient coaxial design network for large systems evolved to consist of cascades of amplifiers of identical gain separated by lengths of cable whose loss just matched the amplifier's gain (although the loss through coaxial cable varies approximately as the square root of frequency, "equalizing networks" were used to make the loss flat across the entire FDM spectrum). Thus, the input and output levels of each amplifier were nominally identical. Under those circumstances, the carrier-to-noise and carrier-to-intermodulation-product ratios degraded as the logarithm of the number of cascaded amplifiers.

In a typical network, a relatively long "trunk" cascade of amplifiers, optimized for noise performance, would branch as required to feed shorter "distribution" legs, which usually contained only two or three cascaded amplifiers operated at relatively higher levels, with the intervening cable spans feeding strings of customer "taps," where a portion of the signals would be diverted to feed individual dwellings. As Figure 1.1 shows, the whole network resembled a tree with branches and sub-branches.

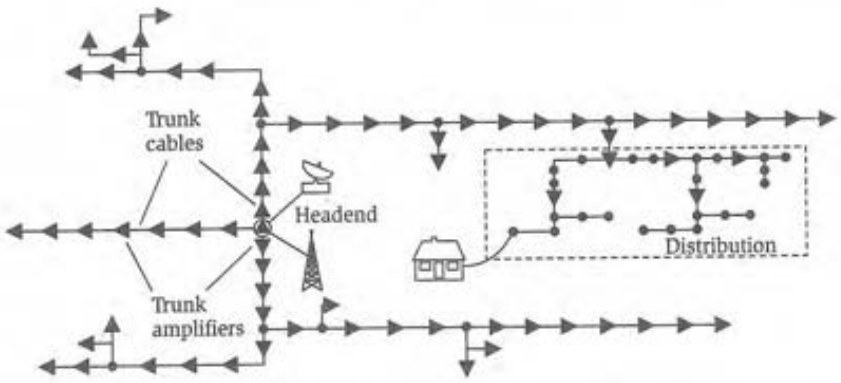


Figure 1.1 Tree-and-branch cable topology.

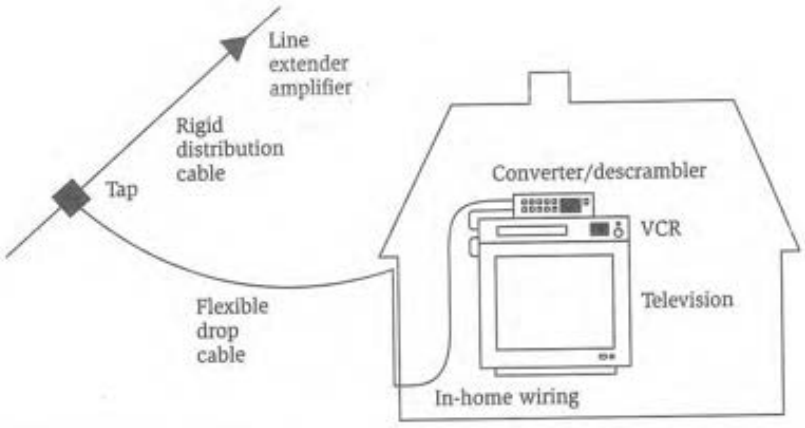


Figure 1.2 Terminal equipment and cable drop.

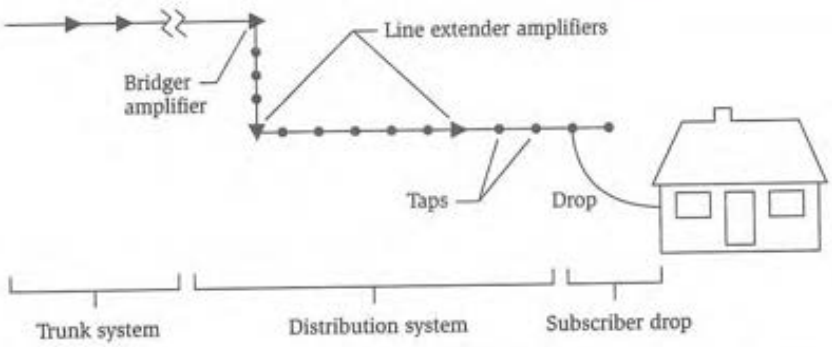


Figure 1.3 Distribution plant.

At each customer's house, a flexible cable was attached to the tap and routed to the terminal equipment, as shown in Figure 1.2. Figure 1.3 distinguishes between trunk, distribution, and drop.

1.5 Coaxial System Limitations

After some experimentation, the optimal design for coaxial trunk lines evolved to strings of amplifiers with gains of approximately 20 dB inter-

spersed with coaxial cables with the same loss at the maximum being carried. In this configuration, each amplifier contributes an equal amount of noise (inversely proportional to input signal level) and an equal amount of intermodulation distortion products (proportional to some power of the output signal level and approximately the square of the number of signals being carried).

The development of cable television distribution technology has been driven primarily by the ever-rising downstream upper frequency limit, which in turn is driven by the need to carry increasing numbers of modulated signals. A secondary need has been for physically larger systems as smaller, formerly independent cable operators have consolidated into large regional operations. In response to these needs, manufacturers have responded with the following.

- Amplifiers with reduced noise figures, and thus reduced per-amplifier noise addition, for a fixed input level
- Amplifiers with increased bandwidths to handle the required frequency range
- Amplifiers with increased power-handling capability, and thus reduced intermodulation distortion, for a given output power level and number of channels
- Cables with reduced loss, both through improved dielectric materials and through the use of larger sizes of conductors with reduced ohmic losses

Though increased gain might be thought of as an improvement, because fewer amplifiers would be required for a given distance, higher-gain amplifiers would have to operate with either reduced input levels (and thus contribute more noise) or higher output levels (and thus contribute higher intermodulation distortion). Conversely, reduced gain per amplifier would require additional amplifiers, and thus drive costs and power consumption higher. A 20-dB net gain (plus or minus a few decibels) has proven a good compromise through many cycles of network evolution.

The possible improvements in component technology are limited. Cables cannot be made arbitrarily large, due to manufacturing and installation costs; amplifiers cannot have higher power-handling capabilities without driving up power consumption and associated heat dissipation; and noise figures are limited by physics and the practical manufacturing costs for devices that are mounted outside and thereby exposed to temperature extremes. The result is that, for a given technology and carried signal complement, there is a maximum possible length of trunk line that will allow any defined level of noise and distortion to be met. Furthermore, if the bandwidth of that trunk line is enlarged and more signals are added, the possible maximum length and amplifier cascade will decrease, because each amplifier will add more noise, distortion, or both.

One possible solution to this dilemma is the use of terrestrial microwave to cover at least a portion of the distance that would otherwise be covered by the trunk cable. The FCC established the cable antenna relay service (CARS) for this purpose. When used for trunking applications, the entire FDM spectrum is up-converted to a microwave frequency at the transmit end and then down-converted to the original frequency range at each receiving site, with the recovered signals inserted into a conventional coaxial distribution network, a technology known as *amplitude modulated link*, or AML.

Such AML trunking typically has noise and distortion performance comparable to about a 10-amplifier coaxial cascade yet can easily cover a distance of 10 or more miles, compared with a few miles for the equivalent coaxial trunk performance. Whether it is lower in cost depends on construction conditions. There are many locations where it is impractical to build coaxial lines but simple to build a microwave link. In other circumstances, multipath AML systems were built, simply to shorten cascades or to serve larger areas than were possible using only coaxial technology.

1.6 The Introduction of Fiber Optics

The major breakthrough in distribution system technology came with the development of practical, linear optical transport. Digitally (on–off) modulated optics had been in use for some time in the telecommunications industry, but in the early 1990s sufficient linearity and optical power levels were developed that it became practical to modulate the light level of distributed feedback (DFB) lasers with the composite FDM spectrum of signals that cable operators wished to transport. As with AML microwave, the performance of a single optical link was approximately equal to that of a 10-amplifier coaxial cascade. However, since optical fibers have a loss that is approximately 1% of the loss of coaxial trunk cable, the opportunity was immediately opened to build larger systems, shorter coaxial cascades, wider bandwidths, or some combination of those parameters.

The earliest application of linear optical transport in trunking was to shorten cascades so that bandwidth could be increased, as shown in Figure 1.4. For several years, the decreasing cost and improving performance of optical technology resulted in ever-smaller node-serving areas, shorter cascades, and increasing bandwidth. Today, off-the-shelf optical transmitters have downstream upper frequency limits of up to 870 MHz and the capability of handling 100+ analog video channels plus other signals with demodulated C/N of about 53 dB and carrier-to-intermodulation-product ratios of 65 dB.

More recent developments have included the development of high-powered transmitters (up to +17 dBm) and the use of the 1550-nm as well as 1310-nm optical wavelengths. Using 1550 nm enables operators to take advantage of fiber's lower loss at that wavelength, the use of optical amplifiers, and access to wavelength division multiplexing (WDM), which allows multiple optical

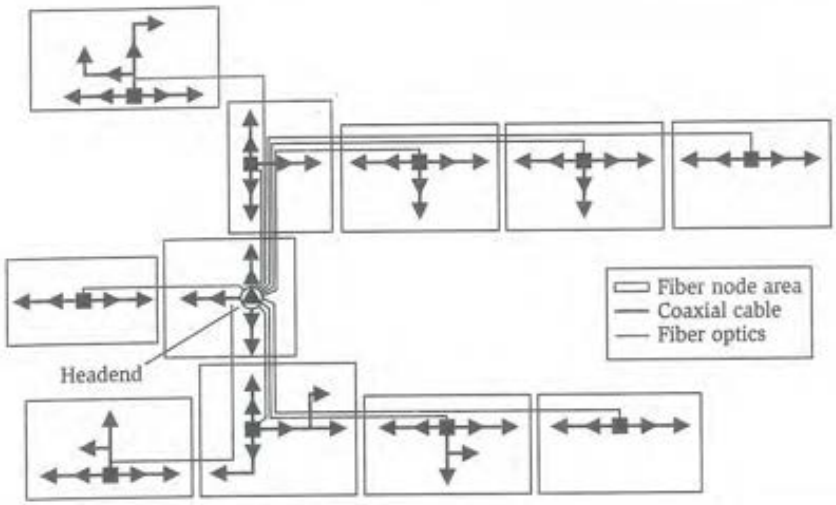


Figure 1.4 Cable system with hybrid fiber-coax topology.

signals to share use of a single fiber. Unamplified links in excess of 60 km and amplified links in excess of 200 km are now possible.

A side effect of the creation of smaller node-serving areas is the ability of operators to deliver an independent set of signals to each node, as discussed earlier. A second effect is that upstream signal transport becomes more durable, because the noise "funneling" from downstream amplifiers is reduced and the probability of ingress from external signals is reduced. These characteristics in turn allow practical delivery of subscriber-specific signals, such as are required for telephone, Internet access, and on-demand video services — all services for which communications reliability combined with the bandwidth per subscriber, rather than gross bandwidth, is of paramount importance.

1.7 High-Level Architecture Changes

Through a continuing process of industry consolidation, individual cable systems have grown steadily larger. This, combined with decreasing node sizes, has made it simply impractical to serve every node directly from the headend in a simple "star" network. Not only is the cost of running multiple, dedicated fibers to each node very high, the huge optical cables that would be required in a major metropolitan system would create major single points of failure.

As a result, operators have developed *distributed* networks, in which at least some of the headend functionality is moved to multiple hubs. Typically, the

transport between headend and each hub is entirely optical and redundant, with route-diverse transport for all critical signals. High levels of multiplexing, whether at the baseband digital level, through wavelength division multiplexing, or a combination of both, reduces the required fiber counts between major facilities. In the largest systems, the core headend signal-processing requirements are sometimes duplicated at two locations, so the entire headend is redundant.

1.8 Video Signal Security and Selective Service Delivery Issues

The creation of multiple marketable levels of television programming required the development of means to selectively control access to individual channels or groups of channels. The two earliest of these were known as *negative traps* and *positive traps*. Negative filters (or traps) were simply band-stop filters that *blocked* individual channels or groups of channels by removing a sufficient amount of a signal or group of signals to render the recovered pictures unwatchable; positive traps were narrow bandstop filters used to *enable* reception of individual channels by removing interfering signals inserted at the headend without removing a significant amount of the sideband information.

Traps, however, had many drawbacks: Negative traps damaged adjacent channels to a greater or lesser degree, and positive traps inevitably removed enough sideband information to noticeably degrade reception. These problems, as well as the inevitable insertion loss, increased as the upper bandwidth of systems increased. Additionally, traps offered only moderate security and could easily be removed by customers, modified, or simply duplicated, as required. Finally, they required physical access to install or remove, so modifying a subscriber's receivable channels required a trip to the home or nearest tap.

The issue of degraded reception was addressed through the development of "descrambling" set-top boxes. Signals to be protected were modified at the headend so as to be unwatchable on standard television receivers. The descrambler reversed the scrambling process for whichever channels were programmed into its local memory. In the United States, the most common technique was to eliminate or modify the synchronization pulses that are a crucial part of a modulated NTSC signal, although more sophisticated techniques (such as shuffling the order in which lines of a picture were transmitted) were employed in European systems.

Although many preprogrammed descrambling converters were deployed, they had the limitation that changes to the list of channels being descrambled still required physical changes to the converter. As a side effect, any box not recovered from an ex-customer represented a potential means by which someone else could access premium programming without paying for it.

A second generation of descrambling converters solved the problem of changes to enabled channels by adding *addressability*. An addressable descrambling converter incorporates the ability to download descrambling information

sent from the headend in real time. This not only reduced the cost of providing services sold on a monthly basis, but enabled the provision of PPV, where access needed to be enabled for just the duration of a single event.

Given market studies that showed that the purchase rates for pay-per-view events were strongly influenced by the ease with which a customer could access them, combined with the cost of staffing customer service centers to handle phoned-in orders, the industry began to deploy boxes that included upstream transmitters so that customers could order programming electronically using their remote controls. This was one of the factors driving the deployment of two-way upgrades to RF amplifiers.

Unfortunately, due to the need to control costs, the scrambling methodologies utilized for securing analog television programming were not particularly sophisticated, nor were the boxes "hardened" against modification, with the result that many were modified to enable reception of more channels than originally intended, and a black market industry in both modified and new "pirate" boxes developed.

The advent of digitally compressed television, as discussed earlier, allowed operators to carry many more channels in the same bandwidth. Due to the quadrature amplitude modulation (QAM) used, the signals were also more durable and could be transmitted through channels with somewhat lower fidelity than analog video channels (though still much greater than for baseband digital signals).

While much more secure, the digital boxes also require significant internal computational ability and digital memory capacity. Thus, they are more expensive than comparable analog addressable converters. Nevertheless, the advantages of digital are such that they are rapidly replacing analog techniques for security and control of video programming. They are almost universally two-way and include both a primary receive channel and a second channel for simultaneous reception of control data from the headend. This has provided a durable and flexible platform for interactive services, such as VOD, interactive gaming, and e-commerce, which require low-latency, high-bandwidth, two-way communications between the headend and individual customers.

1.9 Consumer Equipment Interface Issues

Except to the extent blocked by filters, downstream cable television signals are transmitted throughout customers' homes so that television receivers, set-top boxes, data modems, and other terminal equipment can be connected to any cable outlet. For the same reason, signals generated from anywhere in the home that are within the upstream transmission band (extending roughly from 5 to 40 MHz) will be carried toward the headend. Though this arrangement provides a lot of installation flexibility, it has also created several layers of interface issues for both system operators and consumers.

1.9.1 Shielding Integrity

As discussed earlier, cable television systems transmit signals on frequencies that are used for terrestrial transmission by other entities. This is possible because cable systems are well-shielded closed systems. The weakest link in this shielding is the wiring in customers' homes and the terminal devices connected within the home. By regulation, the wiring within each home is the property of the homeowner and may be installed or modified by them at will. The requirement on the cable operator is to monitor the entire plant for egress and, if owner-controlled wiring is leaking, to work with customers to remedy the problem or, in the ultimate, to disconnect service until it is remedied.

While egress is a legal problem, ingress is a greater threat to system operation. Generally, ingressing signals cause transmission problems before egressing signals exceed legal limits.

Defects in shielding can occur due to the use of poor-quality cable, loose or poorly installed connectors, inadequately shielded components (such as splitters or switches), or poorly shielded television receivers connected to outlets. Unfortunately, the regulations covering the shielding integrity of consumer electronics devices, including television receivers, are not adequate to ensure adequate system egress and ingress performance.

1.9.2 Directly Connected Television Receivers

Almost universally, cable operators deliver their lowest level of video service in analog, unscrambled form, so subscribers need only connect a television receiver with adequate tuning range directly to a broadband outlet in their home. Such "whole-house" reception capability is then marketed as a superior alternative to, for instance, DBS reception, in which a set-top box is required for each television set. In addition to shielding issues, however, connecting television receivers directly to cable outlets raises a number of other interface issues.

Received Signal Quality

Television receivers are designed for connection to antennas. In that environment, a primary requirement is adequate low-signal performance. Generally, fewer than 20 signals are present in the antenna downlead and adjacent channels are not present. By contrast, cable systems deliver roughly 100 channels on contiguous channels at relatively high signal levels. This requires receivers to have better signal-handling characteristics and adjacent-channel rejection filters.

Additionally, devices within the customer's home or the homes of his neighbors may generate upstream signals at levels more than 100,000 times higher than received signals. Due to finite system isolations, these signals will also be present at the input terminals of the receiver, so it must be designed to adequately reject them.

Finally, television receivers generate various internal signals used in signal-processing steps. To the extent that these are present at the input terminals and fall in the upstream band, they degrade two-way system performance. Thus, directly connected receivers need to be designed with high internal isolation.

1.9.3 Functionality Limitations Due to Set-Top Boxes

The imposition of a set-top box between the cable outlet and a subscriber's receiver, while it enables additional services to be delivered, affects functionality. For example, if a subscriber has a video cassette recorder (VCR) connected to the receiver, then the single-tuner set-top box will block the ability to record one channel while viewing another channel. Moving the VCR to a position ahead of the set-top box restores the ability for independent recording but prevents the user from recording any program that is scrambled. Similar difficulties arise when used with receivers capable of picture-in-picture display or when doing timed, multi-channel recording. Ultimately, all these difficulties arise from the fact that the set-top box has its own tuner, which does not operate under the control of the receiver.

1.9.4 Subscriber-Owned Terminal Equipment

Clearly, one answer is for the functionality of the set-top box to move within the receiver. By placing the descrambling process behind a tuner that is under the control of the user, most of the limitations caused by the use of external tuner/descramblers disappear. Despite the efforts described earlier, this is still just a future possibility in the television world. By contrast, a robust standardization effort has made it possible in the world of high-speed data services, with the result that cable modems are low in cost and are becoming readily available in retail electronics outlets.

1.10 Summary

In this chapter we have presented a brief overview of cable television's development and current practice in the United States. Starting with a few comments on cable's history, we have reviewed the structure of the cable industry and described the fundamentals of the technology on which it rides.

The industry began as a relay service for distant over-air broadcast signals and, while in that phase, developed the basic coaxial technologies and components required. The addition of independent sources of satellite-delivered programming changed the nature of the broadband business and drove the development of ever-wider-bandwidth systems and associated development of technology to selectively control access to programming.

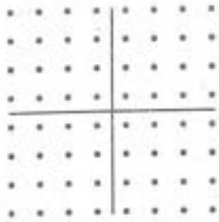
The use of linear fiber-optic technology was driven by a need to expand system size and bandwidth beyond what was possible using all-coaxial

technology, but this brought with it practical two-way transmission and the ability to achieve much higher per-customer bandwidths through the delivery of customized signals to each system node. This in turn enabled the launch of two-way services, such as high-speed Internet access and voice telephony.

Finally, the development of digital video compression has multiplied 10-fold the number of channels that can be transmitted through a given amount of bandwidth. This increased information capacity, combined with the features of digital video set-top boxes, has solved a security problem and also opened the door to the provision of sophisticated interactive services. The rest of this book will fill in the details. Now it's time to roll up our sleeves and dive in!

Endnotes

1. *Cable Television Developments*. Washington, DC: NCTA, summer 1999.
2. For example, draft *EIA Interim Standard IS23* (1994) covered RF characteristics, *EIA Standard EIA-105* (1998) defined an interface that could be used with either analog or digital decoder modules, and *EIA Standard EIA-679* (1998) defined a system whereby digital decryption security could be separated from the receiver.
3. "Time to Plug and Play—Cable, CE Industries Reach TV Accord," *Multichannel News*, December 23, 2002.
4. The early history of CATV and the people who played key roles in its development has been well documented in: Archer Taylor, *History Between Their Ears—Recollections of Pioneer CATV Engineers*, Denver. The Cable Center, 2000.
5. The nontechnical story of the development of HDTV, with reference to the development of digitally compressed video in general, is well told in: Joel Brinkley, *Defining Vision—The Battle for the Future of Television*. New York: Harcourt Brace, 1997.
6. Walter S. Ciciora, *Cable Television in the United States, An Overview*. Louisville, CO: CableLabs, 1995.
7. Bronwen Lindsay Jones and James A. Turner, *Subjective Assessment of Cable Impairments on Television Picture Quality*. Louisville, CO: CableLabs, June 1992.



Part Two
The Signals

Chapter 2
Analog Television

2.1 Introduction

One of the most successful technologies of our time is based upon the standard created by the National Television Systems Committee (NTSC). Like the U.S. Constitution, it seems to have anticipated much of its future and adaptively followed the evolution of its world. Starting with vacuum tubes and progressing to integrated circuits, analog television has survived and prospered. Though much of the current excitement centers on digital television, it is reasonable to assume that analog television will be with us for many more years. Even after digital receivers significantly penetrate households, they will serve the principal viewing room while older analog receivers will be found throughout the rest of the residence.

2.2 Baseband Monochrome Video and Monaural Audio

The technology of analog television is the foundation of the cable industry. An understanding of analog television is key to fully appreciating the rest of the subject matter of this book. Understanding the television signal is a basic requirement for designing systems to deliver it to subscribers.

2.2.1 Television Briefly

The Federal Communications Commission (FCC) adopted standards for commercial television broadcasting in the United States in 1941.[1] It is an analog system, in which the picture is transmitted via vestigial-sideband modulation of the visual carrier and the sound is frequency modulated on a separate carrier. In 1954, the NTSC "compatibly" extended the system to include color information by increasing the utilization of the 6-MHz spectrum occupied by the television channel while allowing continued use of essentially all older receivers with little performance degradation. In 1976, the FCC reserved line 21 for the closed captioning system, and in 1983 a teletext service was authorized. In 1984, the NTSC system was again compatibly extended to include stereo sound. Most

recently, in June 1996, the FCC authorized additional ancillary data services in analog television. These technologies will be considered in this chapter.

The NTSC television system is termed *analog* because the signals representing the picture and sound information can take on *any* value between the minimum and maximum limits. In particular, the strength of the visual signal is inversely related to the picture brightness, with black portions of the picture having the most transmitted power. This negative modulation of the video information was found to reduce the impact of Gaussian white noise on the displayed image. Periodic pulses are included at powers greater than those used to represent black areas in the picture. These pulses provide the timing information required to synchronize the transmitter and the receiver, allowing the picture to be painted correctly on the screen.

Understanding television technology is aided by some appreciation of its history.

2.2.2 Scanning

A picture can be regarded as the composite of many points of light called *picture elements*, or *pixels*. The information for each of these points of light must be derived from the object or scene to be televised, organized in a logical manner, conveyed to the receive site, and then displayed as an image. Since the number of points of light needed to make up a picture is huge, it is not practical to convey to each of them via its own transmission channel. Scanning is a simplifying process that allows a single transmission channel.[2]

The process of scanning was first used on a large scale to transmit still images for use in newspapers (see Figure 2.1). A photograph is placed on a rotating drum and a light-sensitive device is focused on it. As the drum rotates, the light-sensitive device follows a path across the image. With each revolution of the drum, the light-sensitive device is advanced down the length of the rotating drum. This creates a helical scan that follows the path of an imaginary

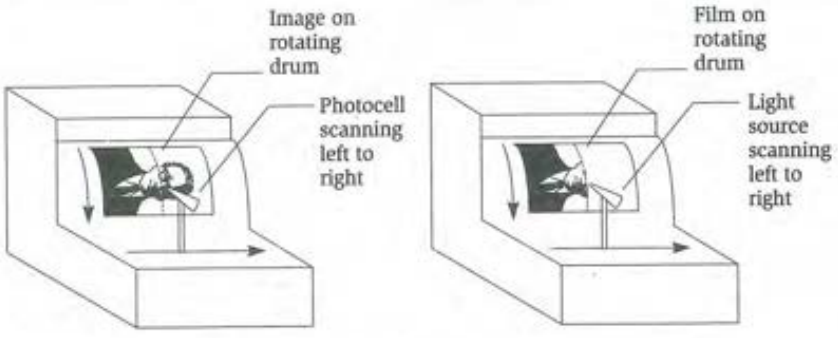


Figure 2.1 Scanning of photograph on rotating drum for transmission to receiver with light source and photosensitive film.

coil spring across the photograph on the scanning drum. The electrical output of the light-sensitive device is proportional to the reflectance of the image, bright areas yielding a larger electrical signal than dark areas. At the receive location, the process is reversed. A light-sensitive film is attached to a rotating drum, and a light source scans the same helical path across the film. High electrical input results in a bright light and a full exposure, or blackening of the film. Conversely, low electrical input yields a lower exposure, or a more transparent area on the film. In this way, a negative image is produced. Standard photographic techniques yield a comprehensible image. It is important that the rotating drums at the transmit site and the receive site be synchronized. They must rotate at the same speed, and the start of scanning must occur simultaneously in both locations. Likewise, the rate of advance of the helical path down the drum must be the same in both locations. If this synchronization does not occur, then geometric distortions are introduced in the reproduced image. Modest distortions involve stretching or compressing of the reproduced image as if it were printed on a rubber sheet and then the rubber sheet distorted. More destructive distortions are possible if the speeds are very different in the two locations.

Although scanning works well when the receive site includes a photosensitive film that can accept and retain the image, the problem becomes a bit more complex when the receiver is the human eye. The human eye exhibits a phenomenon termed *persistence of vision* that allows it to continue "seeing" a light for a brief time after the light is removed. The application of the scanning principle to direct human vision requires that the scanning process be executed repetitively and quickly enough for the eye to fuse all the scan lines into an image.

In 1883, the German engineer Paul Nipkow invented a scanning method consisting of a rotating mechanical disk with appropriately located holes arranged in a spiral (see Figure 2.2). As the disk rotates, only one hole at a time is over the image to be transmitted (see Figure 2.3). As one of the holes completes a left-to-right scan, another hole commences scanning from the left slightly below the path of the previous hole. When the bottom hole completes its scan, the scanning disk will have made a full revolution, and the top of the image is again scanned. Nipkow was granted a German patent but did not actually build the system. In 1923, John Logie Baird of the United Kingdom built a crude system based on the Nipkow scanning disk. It was clumsy and produced poor pictures, but it confirmed the principle of scanning.

Television did not become practical until electronic display tubes and pickup tubes were introduced. In 1923, Vladimir Zworykin, a Russian immigrant who became a U.S. citizen, applied for a patent for a completely electronic television system. The interest in television was intense at that time, and many inventors were active in the field. Space does not allow a discussion of the numerous contributions and the interesting patent fights among inventors. This fascinating story is described elsewhere in detail and is recommended reading.[3]

In 1929, Zworykin applied for a patent that made the monochrome (black-and-white) picture tube practical (see Figure 2.4). An *electron gun* launches electrons

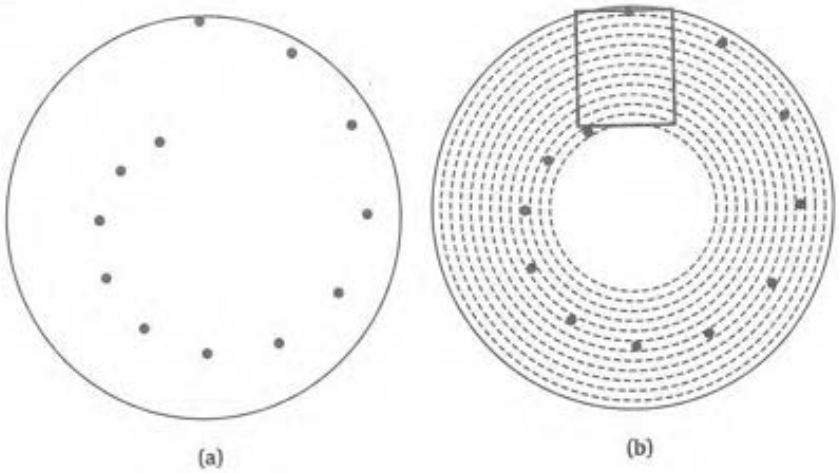


Figure 2.2 Nipkow disk operation. (a) Nipkow disk with a spiral of apertures. (b) Nipkow disk's scanning trajectory through the image plane.

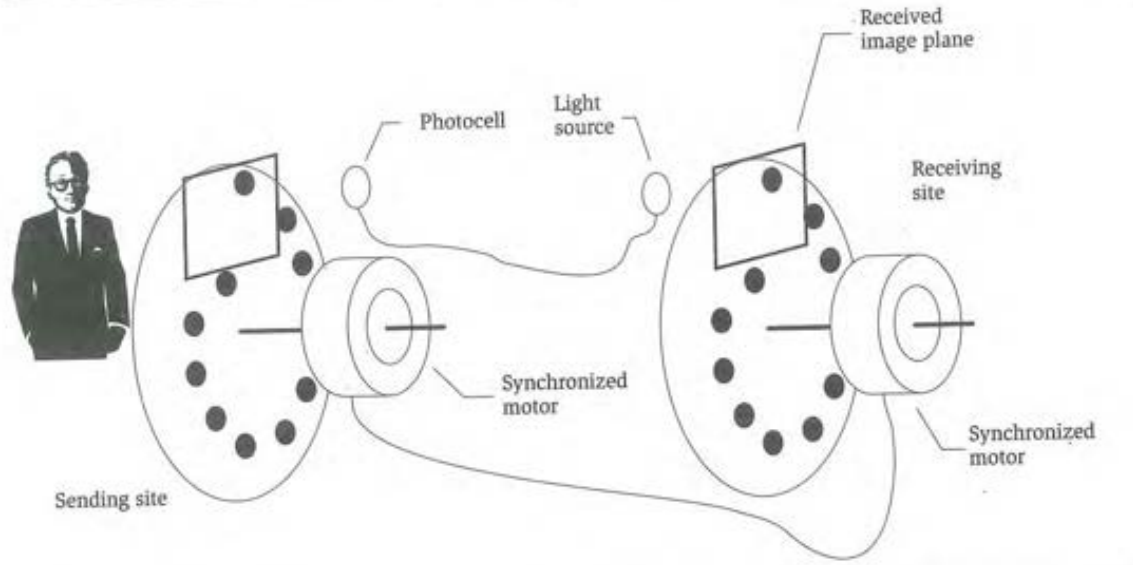


Figure 2.3 Only one hole at a time is exposed to the image on the sending side and the light source on the receiving side.

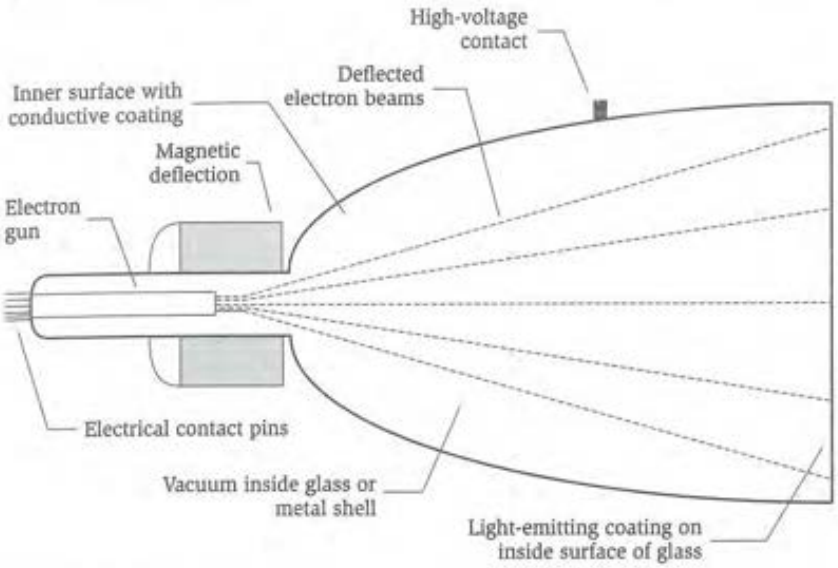


Figure 2.4 Electronic television is based on the kinescope, or cathode ray tube.

past a control grid that regulates the strength of the electron beam. The beam then passes through focusing elements that concentrate the electron beam into a small spot on a glass faceplate that is covered with a fluorescent coating on its inside surface. When the electron beam strikes the fluorescent coating, light is emitted in proportion to the strength of the electron beam. Magnetic coils on the neck of the tube deflect the electron beam vertically and horizontally to implement scanning. The scanned rectangle of light on the face of the picture tube is called a *raster*. The picture tube is also called a *kinescope* because it was meant to display moving or kinetic images. Yet another name for the kinescope is the cathode ray tube (CRT), since the part of the electron gun that emits the electrons is called a *cathode*.

In October 1931, Zworykin applied for a patent on a practical electronic television camera tube that he called the *iconoscope* (see Figure 2.5). As with the kinescope, an electron gun forms an electron beam that is scanned across a surface by magnetic deflection coils. In the case of the iconoscope, the surface is coated with a light-sensitive material. Light focused on the sensitive material causes it to develop an electrical charge in proportion to the strength of the light. When the electron beam strikes the light-sensitive surface, it interacts with the surface in proportion to the accumulated charge. A signal plate on the back side of the light-sensitive surface produces an electrical signal proportional to the light that fell on the surface. Again, many other inventors were working on similar devices, and more patent struggles occurred.

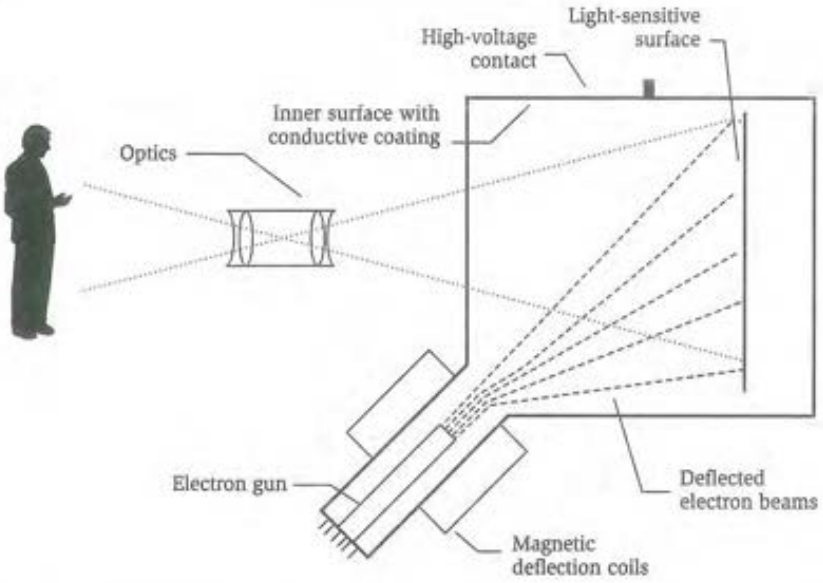


Figure 2.5 Electronic television is also based on Zworykin's iconoscope.

After a scan line is completed, it is necessary to move the electron beam back to the left side of the screen.[4] Ideally, this *retrace* would be done instantaneously, since no picture information is transmitted during the backward motion. Both the scanning drum of Figure 2.1 and the scanning disk of Figure 2.2 did not require retrace. However, electronic deflection is accomplished with magnetic fields, and the magnetic fields store energy. Deflection circuits require time to change the magnitude and direction of the energy stored in the magnetic deflection fields. As a result, it is not possible to instantaneously move the electron beam from the right side of the screen to the left or from the bottom to the top. During the time the electron beam is in retrace, it must be turned off (a process called *blanking*). The period during which the beam is turned off in horizontal retrace is called the *horizontal blanking interval* (HBI), while the much-longer period required to move from bottom to top is the *vertical blanking interval* (VBI). The combination of the two blanking intervals constitutes around 25% of the total time and is the price paid for using cost-effective magnetic deflection. It is interesting to note that modern liquid crystal display (LCD) panels do not require retrace time.

Thomas Edison is credited with developing practical *moving pictures*. Moving pictures depend upon a rapid display of a succession of still pictures. In modern motion picture practice, 24 pictures are each displayed twice per second, but in early movie history the rate was reduced, leading to noticeable "flickers."

The human response to flicker is very complex and is related to the color, size, and brightness of the image. A small increase in the brightness, or a decrease in the repetition rate, results in a much higher degree of noticeable flicker. The minimum rate that is not objectionable is called the *flicker fusion frequency*. A flickering light source at frequencies above that value appears as continuous illumination. For example, fluorescent tubes operated at 60 Hz (and thus extinguished 120 times per second) appear to be continuously "on" because they are operated above the flicker fusion frequency. Monochrome television developers settled on 30 complete images per second in the United States and 25 images per second in Europe. Because of the interlace technique, described later, 60 half-images (or 50 in Europe) are displayed per second. These choices were originally tied to the power line frequency causing any unfiltered power line "hum" that got into the picture to be stationary on the screen. (It was determined that slowly moving patterns of interference were much more objectionable than stationary patterns.) In the European systems, the brightness level at which television flicker is not objectionable is only about 30% of that in the United States. Many viewers accustomed to 60-pictures-per-second television find that the European systems flicker excessively at the brightness levels normally used.

Many of the specifications for the NTSC television system were based on the desire to minimize the cost of receivers using then-current technology and are no longer required. For example, it is now practical to keep the power line hum completely out of video circuits. Consequently, the need to relate image display rate to power line frequency is no longer significant. In Japan, for instance, 50-Hz power is used in the Osaka region and 60-Hz power in the Tokyo area, yet the same television system serves both markets.

The amount of information required to produce pictures of acceptable horizontal and vertical resolution at a 60-pictures-per-second repetition rate is huge. At current resolutions but with a repetition rate of 60 complete pictures per second (that is, without the interlace technique described later), twice the information would be required, and a baseband bandwidth of 8.4 MHz would be required for the video alone. As we shall see, the modulated bandwidth using simple amplitude modulation would be 16.8 MHz — almost three times the current broadcast bandwidth of 6 MHz. Clearly, ways of reducing the bandwidth had to be found that did not impair perceived picture quality.

One of the earliest signal compression techniques involved modifying the scanning system. Sixty pictures per second are required in television because picture tube technology could produce pictures bright enough to watch in daylight with the window drapes open without apparent flicker. Since experience with motion pictures confirmed that 24 complete images per second is an adequate information rate for motion rendition, a technique called *interlaced scanning* was developed whereby half of the scan lines are displayed in one "field" and the other half of the scan lines in another "field." Complete two-field "frames" are transmitted at a rate of 30 per second, providing better motion rendition and flicker frequency than movies yet requiring just 4.2 MHz of baseband bandwidth.

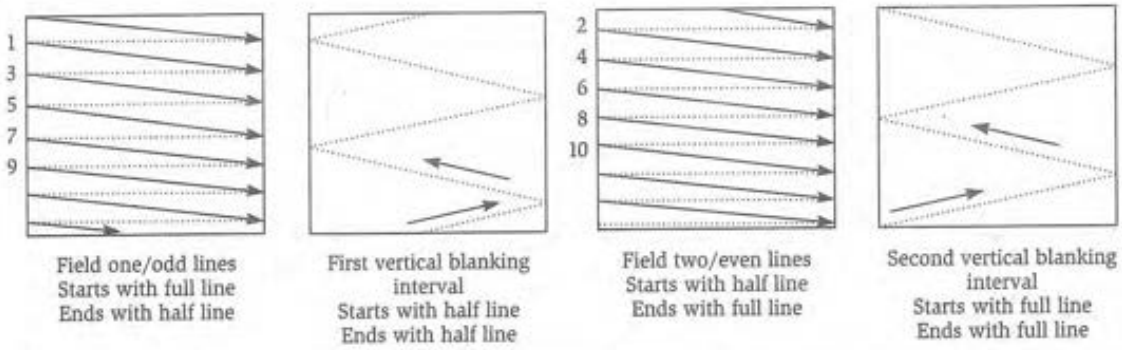


Figure 2.6 Interlaced scanning to reduce flicker while conserving bandwidth.

Interlaced scanning is illustrated in Figure 2.6. A 525-line raster includes two fields of 262.5 lines each. The odd fields start with a whole line and end with a half line, whereas the even fields start with a half line and end with a whole line.

Even though interlace scanning works well with most "natural" video (that is, camera images), interlace scanning can produce inferior results with "synthetic video," which is electronically generated and does not benefit from some of the averaging phenomena of practical cameras. For example, the electron beam in a television camera has a distribution of charge of roughly a Gaussian shape. As a result, it accepts and averages light from adjacent scan lines. This averaging causes a smoothing of the image. Synthetic video, especially when producing text, sharply defines the images of each line. This results in a brightly illuminated edge that is presented at 30 images per second, well below the flicker fusion frequency. This is unacceptable in computer displays that are bright and utilized at very close viewing range. As a consequence, nearly all computer displays use a simple top-to-bottom "progressive scan" technique. Since computer displays are driven by locally generated video, the higher bandwidth requirement of a progressive scan raster is not a problem.

2.2.3 Synchronization

As previously described, there are two fields in each frame of video.[5] Frames have 525 scan lines and repeat 30 times a second, resulting in 63.5 microseconds per horizontal scan line. The reciprocal of this yields the 15,750-Hz monochrome horizontal line frequency. Since fields occur at the rate of about 60 per second, a field lasts 16.67 milliseconds. (All of these values were slightly modified when color was added — see Section 2.4.1)

A horizontal synchronization pulse is required for every line and a vertical synchronization pulse is required for every field in order to display a stable

image. Synchronization (sync) signals are distinguished from the composite signal on the basis of amplitude. The voltage range for visible brightness variations extends from 12.5% to 75% of the unmodulated carrier voltage, leaving 25% for the sync pulses.

Figure 2.7(a) displays an example of a signal that might come from a television camera as it scans one line. At 75% of the maximum voltage, a blanking pulse is inserted, timed to occur during the retrace time. This is shown in Figure 2.7(b). The blanking pulse is nominally 10.9 microseconds, or 16.5% of the horizontal period, while active video occupies 52.6 microseconds. The blanking pulse is at the video signal's black level, keeping the picture tube from emitting light during retrace. The sync pulse is installed on top of the blanking pulse and is at a level referred to as "blacker than black" (see Figure 2.7(c)). The sync pulse is nominally 4.7 microseconds long. Figure 2.8 shows the relative levels of black, blanking, and sync pulses in the modulated envelope of the signal.

Figure 2.9 shows the horizontal sync pulses and video of Figure 2.8 plus additional sync pulses for the vertical blanking interval. Each field's VBI is 21 lines long, leaving 483 lines for active video. The first nine lines of the VBI are devoted to pulses associated with the vertical synchronization process. The fourth, fifth, and sixth lines in each field contain the vertical sync pulse. This pulse is serrated, creating pulse edges properly located for the horizontal synchronization to be maintained during the vertical retrace. The vertical sync pulses are called *broad pulses*, in comparison to the narrow horizontal sync pulses.

Simple resistor–capacitor *sync separator* circuits can be used to separate the vertical pulses from the horizontal pulses.[6] In the lower portion of Figure 2.10(a), capacitor C1 is charged through series resistor R1 and discharges through the input impedance of the following circuits. The narrow horizontal pulses will not build up a significant voltage because the charge bleeds off during the relatively long time between horizontal sync pulses. However, the broad pulses will build up significant charge. This is shown in more detail in Figure 2.10(b). Those who have studied such circuits will recognize this configuration as a simple low-pass filter, also called an *integrator*. A properly adjusted level-sensitive circuit following the resistor–capacitor network will provide a vertical timing pulse. This pulse is used to synchronize a vertical deflection oscillator, which in turn drives the vertical deflection magnetic coils on the neck of the picture tube. Six "equalizing pulses" are inserted before and after the broad pulses to equalize the charge buildup on the capacitor, preserving the precision of the vertical timing. This is necessary because one field ends with a half line and starts with a whole line, whereas the other field has the opposite configuration. Without this precaution, the timing between the two fields will be upset, and the rasters they produce will not be evenly interlaced. The resulting distortion is called *pairing* because adjacent lines from both fields pair up and blur each other's detail while leaving dark lines between them. Although modern digital sync separator circuits do not need the equalizing pulses, the original receivers using vacuum tubes had to rely on tailoring the signal to allow simple circuits.

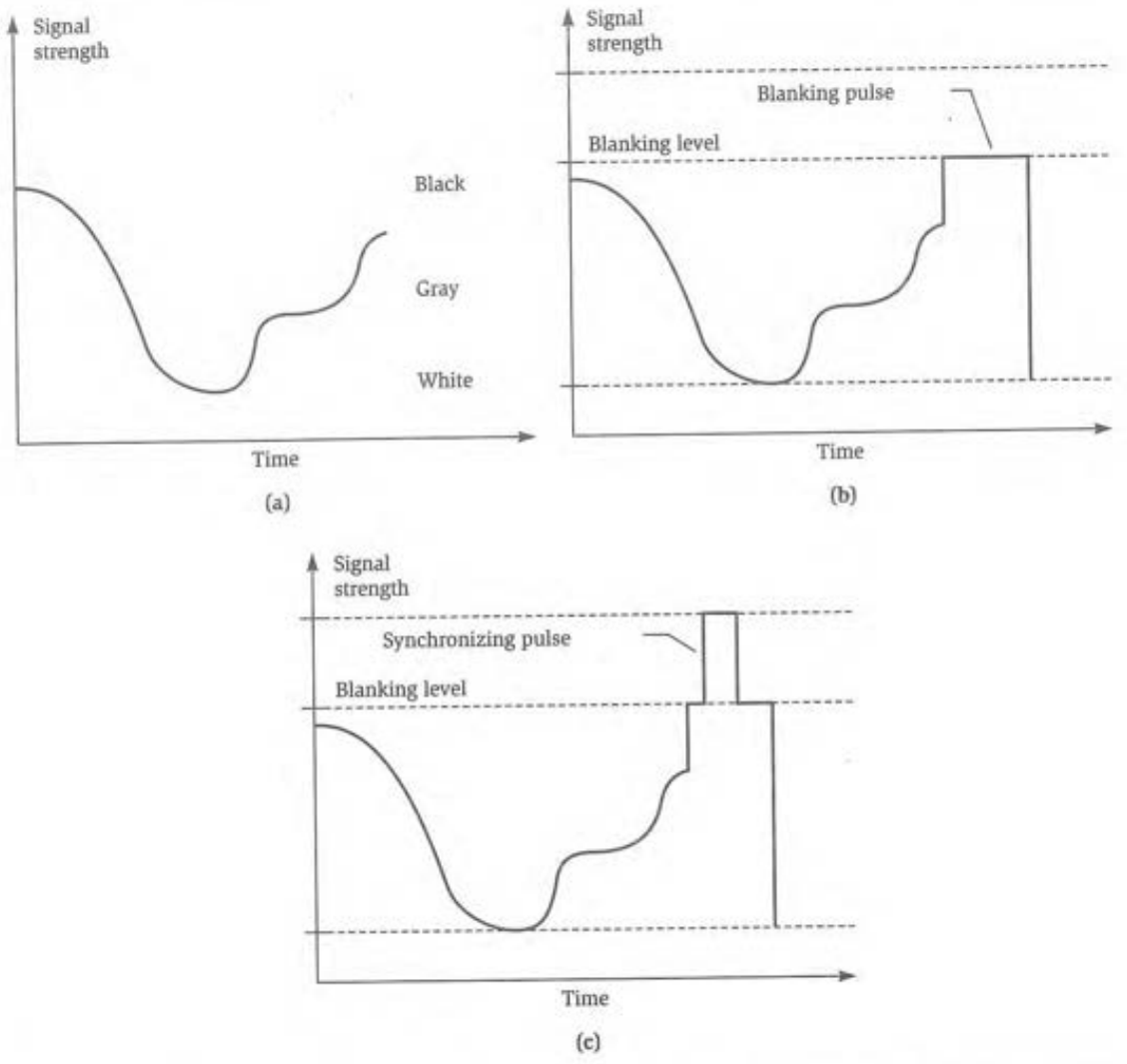


Figure 2.7 Construction of a video signal. (a) Video signal from television camera. (b) Blanking pulse added to avoid diagonal retrace lines. (c) Horizontal synchronization pulse added to facilitate left-to-right timing.

In the upper portion of Figure 2.10(a), a series capacitor C2 with a resistor R2 to ground will convert the edges of the pulses into positive- and negative-going spikes that can be detected by appropriate level-sensing circuits. Note that the serrations in the broad pulses allow a continuous flow of uninterrupted pulses at the correct timing for horizontal synchronization. When these pulses occur

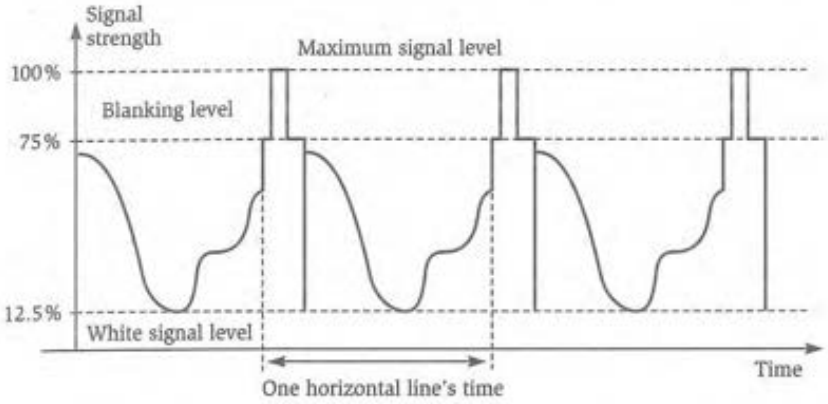


Figure 2.8 Composite monochrome television signal levels (not to scale).

halfway through a horizontal line, the horizontal circuits ignore them. They are effective only when they are near the correct time. Those who have studied such circuits will recognize this configuration as a simple high-pass filter, also called a *differentiator*.

By allocating significant signal resources (about 25% of the signal time and about 25% of the voltage range), rugged synchronization of the video signal was achieved.

2.2.4 Spectrum of Scanned Images

It is well understood that any repetitive signal can be decomposed into a sum of sine and cosine waves of appropriate amplitude, frequency, and phase.[7] There

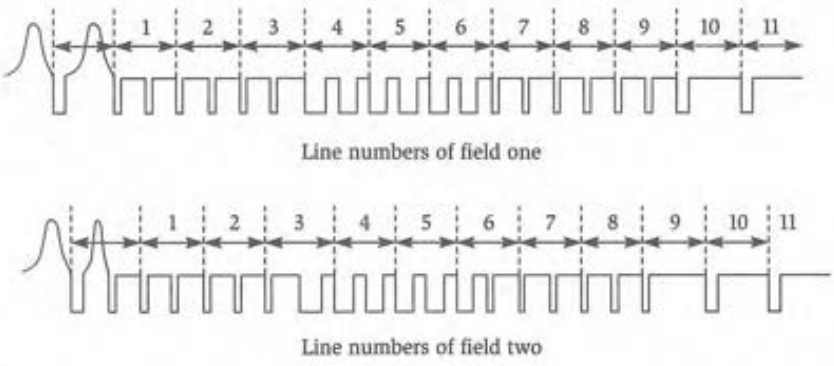


Figure 2.9 Vertical and horizontal synchronizing pulses of field one and field two. Note locations of half lines (not to scale).

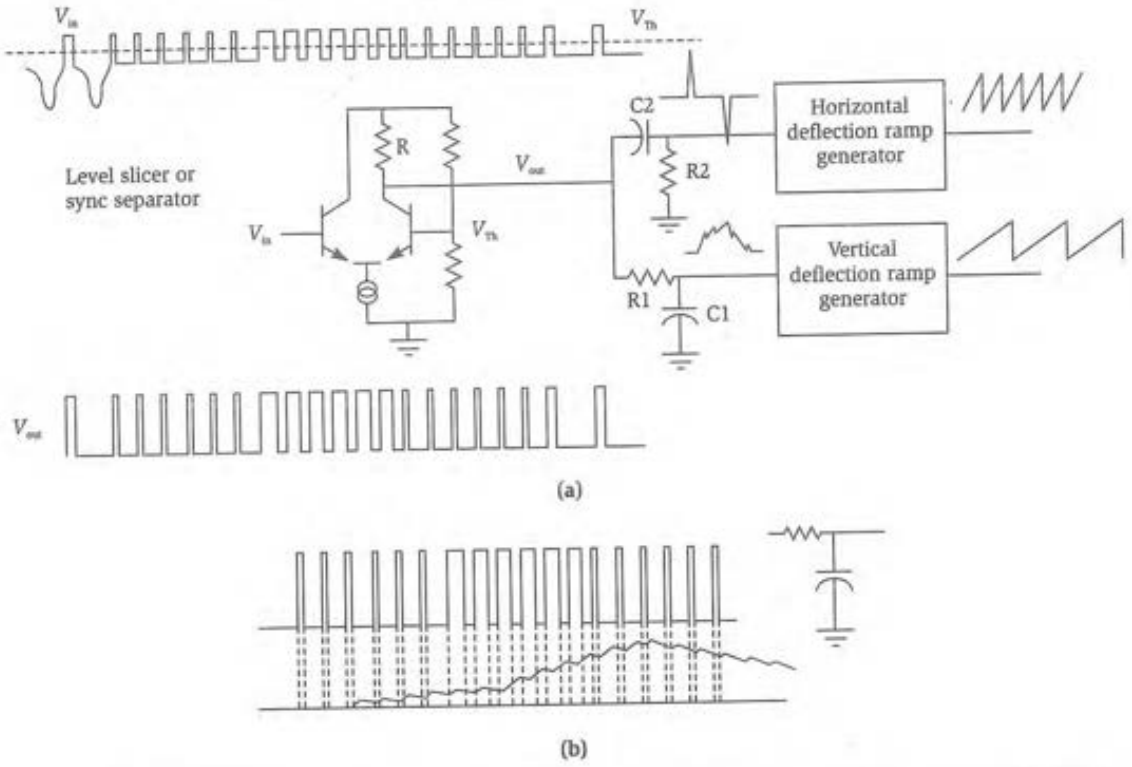


Figure 2.10 Receiver synchronization. (a) Sync separation and deflection ramp generation. (b) Vertical synchronization pulse generation.

are at least three ways we know this: (1) mathematical proof, (2) geometric construction, and (3) measurement with spectrum analyzers. The most rigorous method is the mathematical. Equations (2.1), (2.2), and (2.3) display the familiar Fourier series equations.

$$f(t) = a_0 + \sum_{n=1}^{\infty} [a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)] \qquad (t_0 < t < t_0 + T) \qquad (2.1)$$

Equation (2.1) is the trigonometric Fourier series representation of $f(t)$ over an interval $(t_0, t_0 + T)$.
The various constants a_n and b_n are given by

$$a_n = \frac{\int_{t_0}^{(t_0+T)} f(t) \cos(n\omega_0 t)\, dt}{\int_{t_0}^{(t_0+T)} \cos^2(n\omega_0 t)\, dt} \qquad (2.2)$$

and

$$b_n = \frac{\int_{t_0}^{(t_0+T)} f(t)\sin(n\omega_0 t)dt}{\int_{t_0}^{(t_0+T)} \sin^2(n\omega_0 t)dt} \tag{2.3}$$

We could derive these equations and apply them to scanned images, but that would duplicate what is already available in many other places. In particular, a paper published by Mertz and Gray in the *Bell System Technical Journal* in July 1934, titled "Theory of Scanning and Its Relation to the Transmitted Signal in Telephotography and Television," showed that periodic scanning of an image yields a spectrum with energy clustered around the horizontal scanning frequency, with large gaps in between. This pioneering work was done long before the availability of modern computers. Donald Fink did follow-up work and the results are included in the 1957 *Television Engineering Handbook*.[8] His work clearly showed that the picture's energy is clustered around harmonics of the horizontal scanning frequency, as shown in Figure 2.11.

2.2.5 Television Audio

Television sound is limited to 15 kHz because of the presence of strong spurious signals related to the monochrome horizontal scanning frequency of 15,750 Hz.

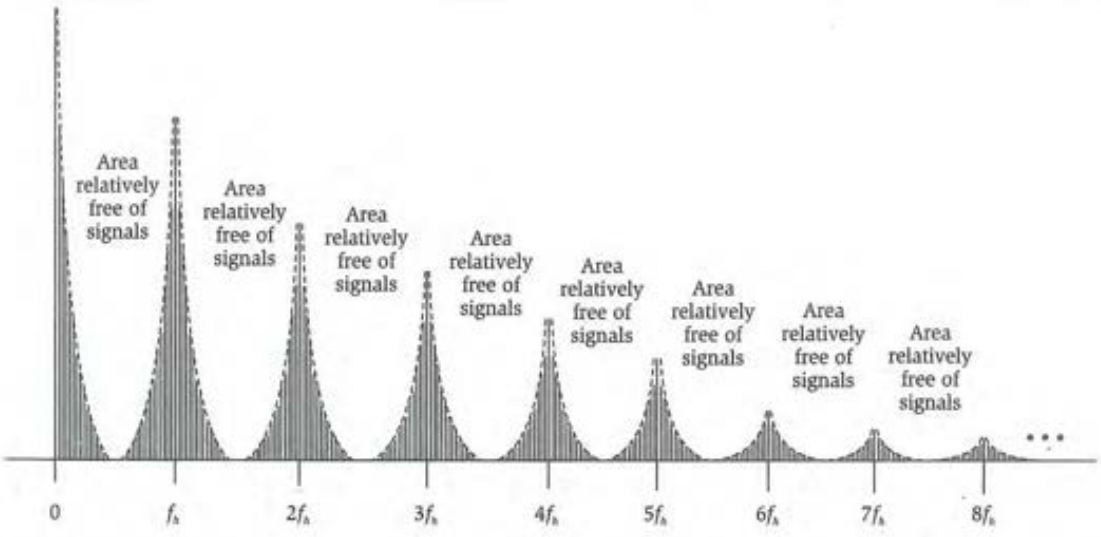


Figure 2.11 Video spectral energy clusters around harmonics of horizontal scan frequency, f_h, all the way up to the upper band edge.

For many years, audio in television was underappreciated. Only in the last decade has it received the attention it deserves as a fundamental part of the viewing experience. Perception experiments with human subjects at the Media Laboratory of the Massachusetts Institute of Technology have demonstrated that subjects shown two identical video presentations with differing accompanying sound qualities believed that the video with the better sound also had better video quality.

2.3 Modulation

Modulation is the process by which an information-bearing baseband signal is impressed upon a "carrier" signal. Baseband frequencies are those from 0 Hz to the maximum frequency content of the signals that come out of a television camera or a microphone. Only one baseband signal can be conveyed per transmission line, since a second signal, occupying the same frequency range, would interfere. Modulation allows multiple baseband signals to occupy different frequency bands in the cable through use of different carrier frequencies. The separately modulated signals are then added together to provide one composite signal for transmission. This process, called *frequency division multiplexing* (FDM), gives rise to the familiar concept of "channels." Each channel is a band of frequencies associated with one video and audio signal.

There are several ways of approaching the subject of amplitude modulation. A phenomenological approach describes the physical processes and follows the historical path by which amplitude modulation was invented and deployed. A mathematical approach starts from first principles and develops relatively simple equations describing amplitude modulation and detection and the special versions of each developed for NTSC television. This approach builds a firm foundation upon which to consider the addition of compatible color to the monochrome television system. Additionally, Chapter 8 explains the modulation process in terms of vector diagrams called *phasors*.

2.3.1 A Phenomenological Approach to Amplitude Modulation

A television camera produces signals that are proportional to the strength of the light in the image and includes synchronization pulses. Figure 2.8 shows the waveform of such a signal. This is an information-bearing baseband signal because the frequencies of this signal extend from zero to the maximum produced by the television camera. The zero-frequency component represents the average brightness of the image. The highest frequency conveys the finest detail of the picture.

A natural method of conveying the television camera signals would be to modulate the strength of a carrier wave in proportion to the amplitude of the original baseband signal, as shown in Figure 2.12.[9] The outline of the maximum

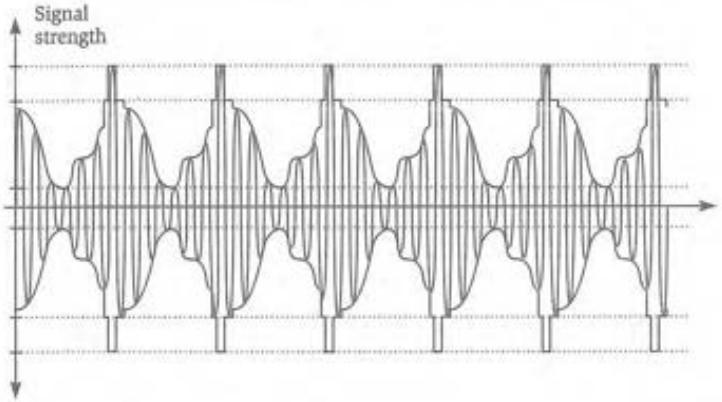


Figure 2.12 Television modulation. Carrier wave (shown at much lower frequency for convenience of illustration) modulated with video signal.

excursions of the carrier, termed the *envelope*, is seen to be a replica of the original baseband information-bearing signal.

At the receive site, the goal is to display the signal from one television camera. This can be done first with a tuned circuit that passes only the desired modulated carrier (actual receivers implement a more sophisticated version of this method based on the superheterodyne receiver principle), followed by a simple diode that will pass only the positive-going portion of the modulated signal, followed by a low-pass filter to remove the carrier pulses, as shown in Figure 2.13.

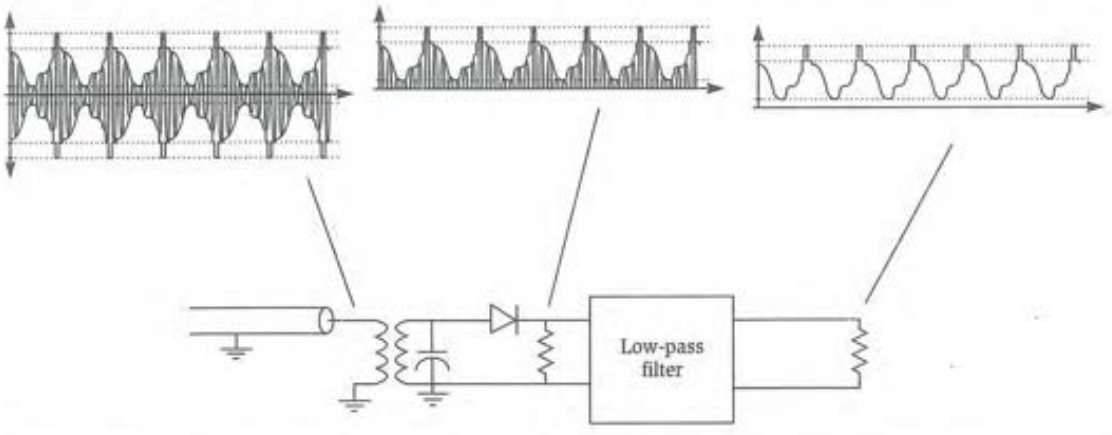


Figure 2.13 Envelope detection with a diode and low-pass filter.

2.3.2 A Mathematical Approach to Amplitude Modulation

Amplitude modulation is the process by which a high-frequency carrier wave is multiplied by a low-frequency information-bearing signal to generate a new signal, centered on the carrier frequency, which includes *sidebands* that carry the information. Through use of multiple carriers, each modulated by a different information-bearing signal, it is possible for multiple information streams to share a common transmission medium.

More formally, let us define the two signals as follows:

$$\begin{aligned} \text{Carrier} &= A\sin(\omega_c t) \\ \text{Information} &= (1 + B\sin(\omega_v t)) \end{aligned} \tag{2.4}$$

where:

Carrier = carrier wave

A = peak amplitude of the carrier wave

ω_c = angular frequency of the carrier wave = $2\pi f_c$

f_c = frequency of the carrier wave, in hertz

Information = information-carrying signal (simplified to a simple sine wave to simplify the equations)

B = peak amplitude of the information-carrying signals (must be less than 1)

ω_v = angular frequency of the video information

t = time, in seconds

If we multiply these, we get

$$A[\sin(\omega_c t) + B\sin(\omega_c t)\sin(\omega_v t)] \tag{2.5}$$

That is, we get the original carrier plus the effect of multiplying the original carrier times the information-carrying signal. When we multiply two sine waves, however, we get two signals whose frequencies are equal to the sum and difference of the two original frequencies and whose amplitudes (voltage, in this case) are in accordance with the standard trigonometric identity:

$$\sin A \sin B = 0.5\cos(A - B) - 0.5\cos(A + B) \tag{2.6}$$

In other words, we get a sideband on each side of the carrier that is spaced from the carrier by the information-carrying frequency and whose amplitude will be $6 + 20 \log$ (B) dB less than the amplitude of the unmodulated carrier.

In the case where the information-carrying signal is a scanned television picture, it will have frequency components extending from zero hertz to an upper frequency limit that is related to the possible horizontal resolution of the displayed image. In the NTSC system, the information-carrying signal is limited to about 4.2 MHz of bandwidth. Thus, with the simple *double sideband* modulation just described, the modulated signal bandwidth would be 8.4 MHz wide just for video information, with the audio still to be fitted into the spectrum.

It is apparent, however, that double sideband modulation is wasteful of spectrum, because each sideband carries exactly the same information. An alternative is *single sideband* (SSB) modulation, whereby only one of the sidebands and no carrier is transmitted (saving both bandwidth and transmitter power). However, the detection process is more complex for SSB signals because a local carrier must be generated to mix with the incoming sideband in order to recover the original information-carrying signal. While this is simple today, it wasn't when the NTSC system was being developed and so was rejected.

The compromise reached was vestigial sideband (VSB) modulation, in which a normally modulated double sideband signal is passed through a filter that passes one sideband (the upper sideband of the transmitted signal), but only 0.75 MHz of the other sideband, as shown in Figure 2.14(a). As can be seen, the sound signal has been inserted 4.5 MHz above the visual carrier, leading to a total visual + audio signal that is only 6 MHz wide yet includes picture information extending up to between 4.0 and 4.2 MHz.

If this signal is demodulated using a standard amplitude detector, the result will be that modulating frequency components below 0.75 MHz will be recovered from both sidebands, while those falling above 1.25 MHz will be recovered from only a single sideband, thereby distorting the relative amplitudes of the detected information signal, as shown in Figure 2.14(b). In order to correct for this, the modulated signal must be passed though a compensating filter before detection. The shape of the required filter is shown in Figure 2.14(c).

2.3.3 Video Bandwidth

When the NTSC system was being designed, a tremendous number of human visual experiments were undertaken. Estimates were made on the likely course of technology. It is a tribute to the work done that the NTSC system has lasted for more than 50 years and is likely to be in use for at least another decade or two.

NTSC television is meant to be a "five times picture height" system. That is, the compromises made between video quality and economics were such that they would not be noticed if the picture were viewed from a distance of at least five times the picture height. Consideration of this and of the bandwidth available, as well as of the ways of occupying that bandwidth in a manner that would allow affordable receivers, led to the choice of 525 total (483 active) scan lines. If a synthetic video pattern is generated by electronically making alternate lines

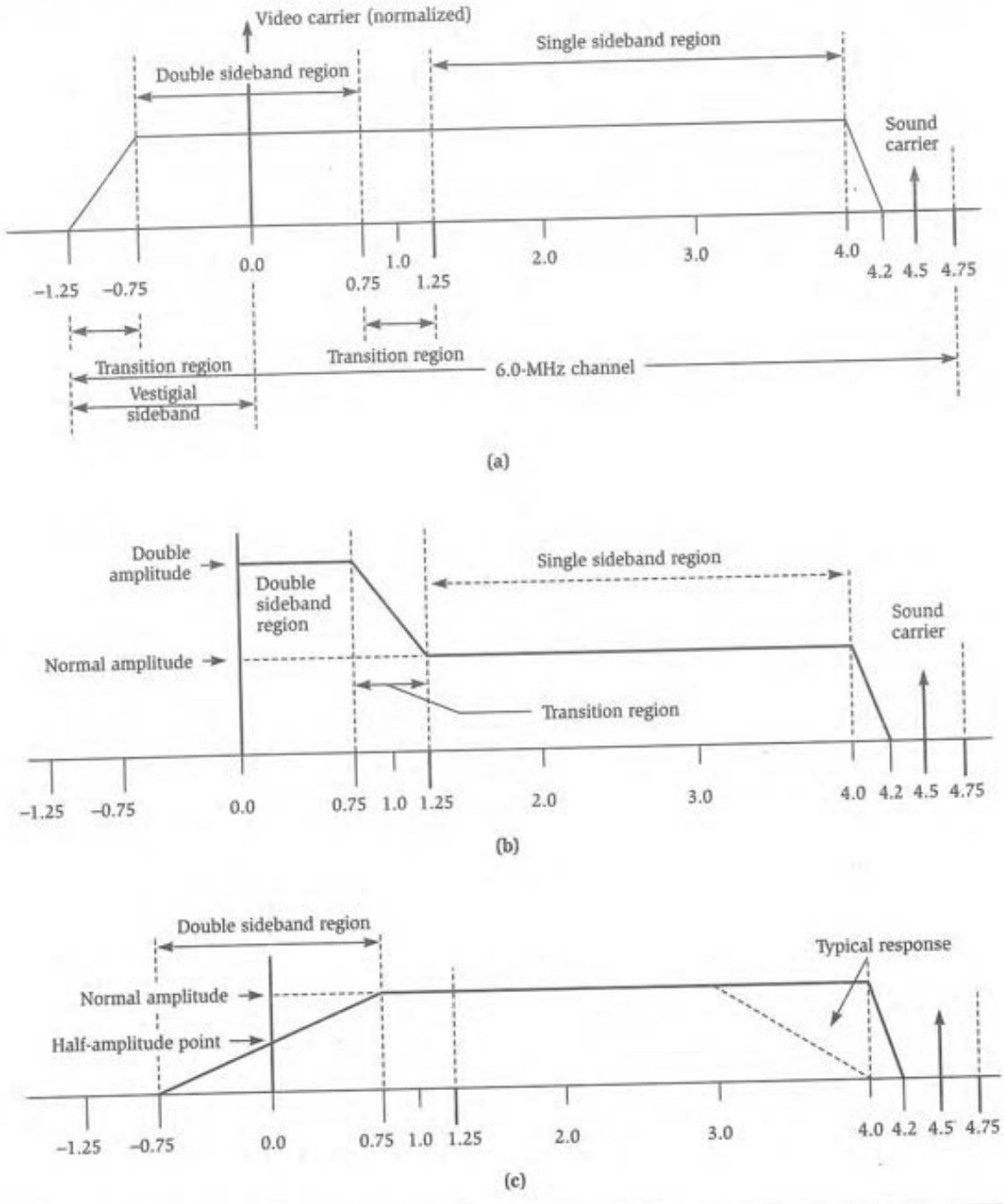


Figure 2.14 VSB reception. (a) Vestigial sideband spectrum normalized to 0.0 Hz as seen at RF (at IF, spectrum is reversed). (b) Output of a "flat response" television receiver as seen at RF (at IF, spectrum is reversed). (c) Idealized and typical TV receiver response curves as seen at RF (at IF, spectrum is reversed).

black and white, and if the whole raster is placed in the viewing area of the screen, then 483 picture elements can theoretically be resolved in the vertical direction. In the "real world," a large number of factors limit vertical resolution, including that the scanning system is a sampling system and that raster lines do not fall exactly on object boundaries. Rather than a clean edge, some of the light falls into two or more adjacent lines, blurring the image. Additionally, there is raster instability. Minor movements between scans add further blurring. The scanning spot in the pickup device and the scanning spot in the display device are not precise points but have a bell-shaped (Gaussian curve) distribution of effectiveness. All these complex factors yield less actual vertical resolution than the number of scan lines would indicate. The ratio between the number of scan lines and the actual vertical resolution varied from 0.5 to 0.85 in various experiments and is commonly accepted as 0.7. R. D. Kell is credited with much of the work on this phenomenon, so this ratio is called the *Kell factor*. That is, 483 active scan lines yield an effective vertical resolution of about 338 lines.

There is no single derivation of the bandwidth needed to carry video. The best that can be done is to give a plausibility argument for why something like 4.2 MHz is a reasonable number. Assuming a desire to have approximately the same horizontal resolution as vertical resolution and noting that the *aspect ratio* (the ratio of vertical height to horizontal width) of an NTSC picture is 3 by 4, we must multiply 338 by 4/3, yielding 450 horizontal lines of resolution. Dividing the aspect-ratio-adjusted number of horizontal picture elements by 2 yields the maximum number of cycles of video in the active horizontal video time. This yields 225 cycles. The 225 cycles must take place in 52.6 microseconds, yielding a frequency of 4.28 MHz.

The maximum number of picture elements possible in a video frame constructed electronically is 483 scan lines times 450 horizontal lines of resolution, or 217,350 pixels. European television systems that use the 50-pictures-per-second rate have more time per picture, allowing for 625 scan lines. To retain a balance between vertical and horizontal resolution, these systems have allocated more bandwidth per channel.

2.3.4 Frequency Modulation

The previous discussion of modulation considered only modifying the strength of a carrier signal to impress an information-bearing baseband signal upon it.[10] A cosine wave, however, has three parameters that can be varied to convey information. Equation (2.7) shows them:

$$s(t) = A \cdot \cos(2\pi ft + \theta) \tag{2.7}$$

Here A is the amplitude, f is the frequency, and θ is the phase angle of the cosine function. The symbol $s(t)$ is the signal as a function of time. As previously discussed, visual information is added to the carrier by varying its amplitude.

Sound information is modulated on a separate carrier by varying its frequency, in which case the signal can be expressed by

$$s(t) = A \cdot \cos(2\pi[f_1 + g(t)]t + \theta) \tag{2.8}$$

where the information-bearing signal is $g(t)$. Without loss of generality, the constant phase angle θ can be dropped, yielding

$$s(t) = A \cdot \cos(2\pi[f_1 + g(t)]t) \tag{2.9}$$

If the baseband information-bearing signal is a simple cosine wave of frequency f_2, the resulting expression is

$$s(t) = A \cdot \cos(2\pi[f_1 + \cos(2\pi f_2 t)]t) \tag{2.10}$$

The mathematics of cosines within cosines rapidly becomes very complex, involving advanced functions and concepts. There are excellent treatments of these details for anyone who is interested. The principle result for our purposes, however, is that the bandwidth of a frequency modulated signal can exceed that of an amplitude modulated signal given the same information-carrying signal. The signal, however, is more robust and can survive noise and other distortions better than amplitude modulation. Frequency modulation (FM) belongs to a class of signals called *spread spectrum* because it spreads the information signal over more bandwidth than it occupies in its baseband version. This technique trades off spectrum inefficiency for signal robustness.

The phenomenological concept motivating frequency modulation is the notion that noise, particularly static, was believed to be only an amplitude effect. If the amplitude of the signal did not carry information, the impact of noise added to the signal would be negligible. This simplistic model is mostly correct, and the main benefits of FM include much less susceptibility to noise.

The creators of NTSC considered it to be a video enhancement to radio. The intention was that, at the outer limits of reception, the audio signal would survive even if video failed,[11] and thus it was made extremely robust by using frequency modulation. In cable practice, we are able to take advantage of this by reducing the relative strength of the audio signal and thus reducing the overall signal loading on the cable system. In NTSC television, the sound signal is frequency modulated onto a carrier that is 4.5 MHz above the visual signal. Over-air broadcasters use separate transmitters for visual and aural signals and combine the signals before feeding a shared antenna.

2.4 Modern Analog Television

Modern analog television differs from the original system in only three respects: the addition of color, stereo sound, and ancillary signals. All this has been accomplished without making earlier products obsolete. If the very first commercial television receiver were operational and powered today, it would still receive and display television programming. "Compatibility," however, is an ambiguous term, and older receivers experience slightly reduced performance. These compromises stand the cost–benefit test and have provided the viewing public with a net positive benefit.

2.4.1 Color Television

Color is the most significant addition to NTSC.[12,13] A large number of trade-offs were necessary in the original NTSC system to make it practical and affordable, but even more were required to add color. The color system carefully matches its design to the human visual system to minimize the incremental information rate required.

The most important discovery is that the human visual system responds to color in two distinct ways: (1) as a response to the frequency (or wavelength) of the light, and (2) as a response to combinations of primary colors. If we pass sunlight through a glass prism, we decompose the light into individual wavelengths that we see as many separate colors. If this were the only way in which the eye perceives color, a color television system, or even color printing and photography, would be all but impossible. Fortunately, the eye, in response to various ratios of red, green, and blue,[14] perceives nearly all the possible colors.

The science of colorimetry is fascinating and complex and well beyond the scope of this brief chapter. Only a few basics can be covered here. In the color receiver, the picture tube is actually three picture tubes in one.[15] There are three electron guns and three sets of colored phosphor elements on the screen. A *shadow mask* in the tube is carefully aligned to allow only the electrons from a particular gun to strike the phosphors of its assigned color. The shadow mask is also called an *aperture mask* because of its array of holes, or apertures (see Figure 2.15). The three electron guns are tilted at an angle with respect to the axis of the picture tube. The three electron beams pass through the same holes in the shadow mask, which is located very close to the screen. Because they aim from different angles in the neck of the tube, the beams land on adjacent but different spots. The spots have phosphors that emit red, green, and blue light. The spots are small and close together, and the eye fuses them into a continuous color from normal viewing distances. The reader is invited to take a magnifying glass and look closely at the screen of a color receiver. Clusters of dots or clusters of small bars of the three different colors will be visible when the screen is illuminated.

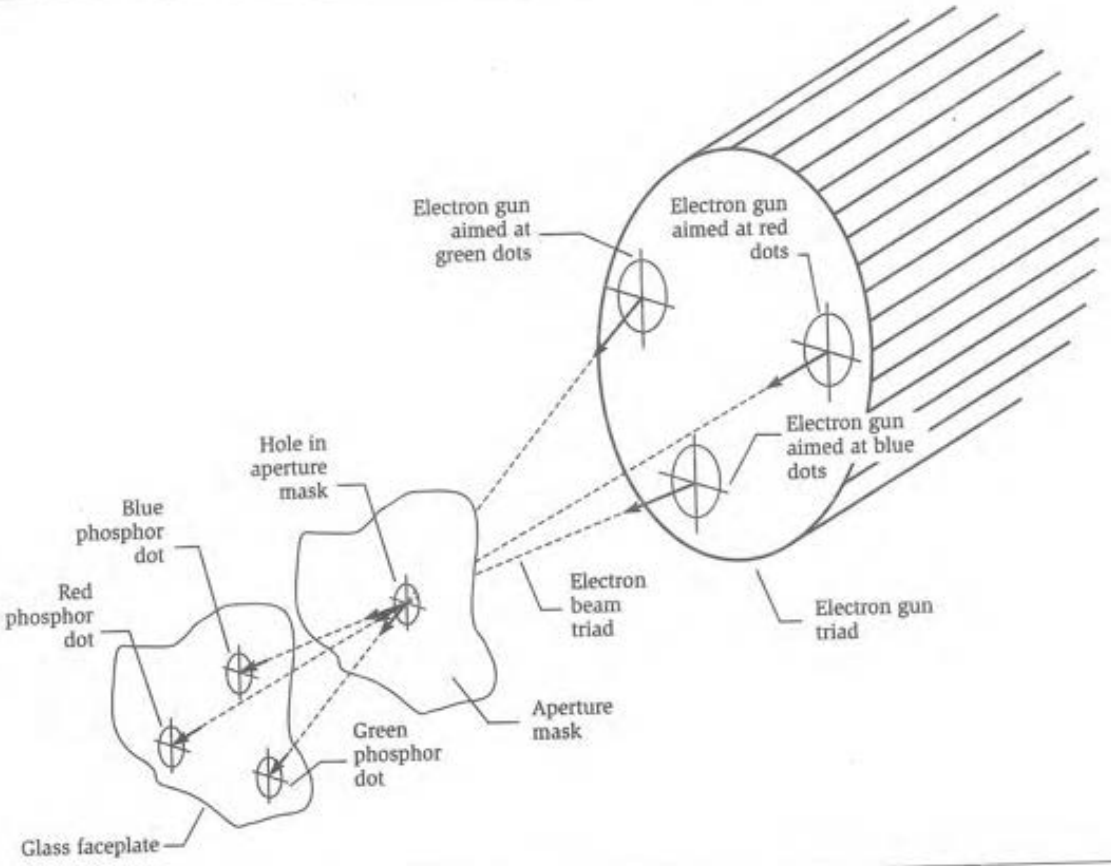


Figure 2.15 Color picture tube structure with colored phosphor dots, aperture (or shadow) mask, and cluster of three electron guns.

If a combination of three complete images in red, green, and blue were required to make a color picture, the required transmission bandwidth would be tripled. Even the improvement color brings might not be worth that drastic a loss of channel capacity. Fortunately, the eye is quite pleased with a "colored" television system and does not require a true "color" system. If we consider the Sunday color comics, we realize that we have a detailed black-and-white image with some color thrown on it. The eye and the imagination do the rest. In colored television, the same approach is taken. The detail is in the luminance (brightness) signal. The color signal contains much lower resolution and only approximately fills in the monochrome outlines. That this is the case can be seen when colored text is in the picture. For example, a yellow letter T on a blue background will have a yellow top bar because the image is large enough

horizontally to be represented by the frequencies involved. However, the vertical component of the T will usually not be yellow. To represent it in yellow would require frequency components beyond what is available.

To make the color signal compatible with older monochrome receivers, it was necessary to combine the signals in a manner that yielded the luminance signal plus two narrower bandwidth "color difference" signals. Since the eye is more sensitive to detail in certain colors than in others, it is possible to reduce the bandwidth of some of the colors more aggressively. The color television system manipulates the color signals using a set of linear matrix equations. This separates the signal into a component that contains flesh tones and near flesh tones and another component that has quite different colors. The principal color signal, termed the *I* signal, has a transmitted bandwidth of around 1.5 MHz. The secondary color signal, termed the *Q* signal, has a transmitted bandwidth of around 0.5 MHz. In a color receiver, linear networks can then separate the signals into the three primary colors, red, green, and blue. Simultaneously, a monochrome receiver will respond only to the luminance signal and ignore the color signals.

Although the NTSC system uses bandwidths of 1.5 MHz in the *I* channel and 0.5 MHz in the *Q* channel, practical color receivers actually implement a bandwidth of only about 0.5 MHz (or less) in each channel. This compares with several megahertz for the luminance signal. The main reason for the reduced bandwidth is the problem of "chroma noise." When the higher-frequency chroma signals are synchronously demodulated (a frequency conversion process), the noise added at those frequencies is also brought down in frequency. This causes the interference resulting from the noise to increase in size on the screen. Reducing the chroma bandwidth minimizes this unpleasant effect, but at a reduction in color resolution.

Even with only 1.5 MHz of *I* signal and 0.5 of *Q* signal, we still need to find a place to put 2.0 MHz of additional information. The principle of compatibility precludes simply adding more bandwidth. This is where the work of Mertz on the spectrum of scanned images provides an answer. Figure 2.11 showed that the signal energy is clustered around multiples of the horizontal scan frequency, with significant spectral gaps. Since this is a consequence of scanning and the color information likewise is a product of scanning, it too has spectral gaps. The trick then is to arrange the two spectra such that the energy clusters interleave. Figure 2.16 shows that if the color signal is modulated onto a carrier that is an odd harmonic of half of the line frequency, the color signal spectrum will interleave with the monochrome signal spectrum. The color signal occupies frequencies that are nearly empty in a monochrome signal.

This technique for hiding the chroma signal in the luma can be visualized in the time domain as well (See Figure 2.17). A cosine wave that is an "odd harmonic of half of the line frequency" will start with the opposite phase on adjacent lines. In the basic case, the first harmonic of half the line frequency will make only half of a cosine wave in the time of one horizontal scan. It will begin

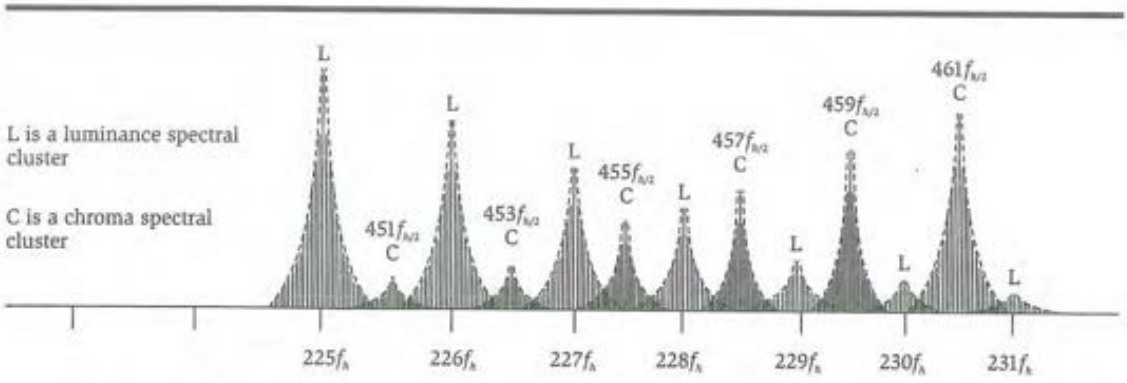


Figure 2.16 Luminance spectral energy clusters around harmonics of horizontal scan frequency f_h all the way up to the upper band edge. Chroma spectra energy components cluster around odd multiples of half horizontal scan frequency and are interleaved with luminance.

the next line with the opposite phase. So where the signal is brightest on one line, it will be dimmest on the adjacent line, and vice versa. These opposite brightnesses on adjacent lines will tend to average out and reduce visibility of the color carrier. The same is true of higher odd harmonics (even harmonics would cause the signal to have the same phase on adjacent lines, and the bright and dark spots would reinforce each other rather than tend to cancel). At the color subcarrier frequency, the bright and dark spots on the screen are very small and are positioned on adjacent lines relative to each other so as to average out the brightness they cause.

The two color signals, I and Q, are modulated in quadrature to each other onto the color subcarrier.[16] *Quadrature modulation* means that two carriers of the same frequency are used, one phase shifted by 90° from the other. The color signal can be expressed as

$$c(t) = I(t) \cdot \cos(2\pi f_c t) + Q(t) \cdot \sin(2\pi f_c t) \tag{2.11}$$

The sine function is 90° phase shifted from the cosine function. The required detection process is called *quadrature detection*. In the receiver, quadrature detection cleanly separates the two signals. This separation is based on the following simple trigonometric identities:

$$\cos(2\pi f_1 t) \cdot \cos(2\pi f_2 t) = \frac{1}{2}\cos(2\pi[f_1 - f_2]t) + \frac{1}{2}\cos(2\pi[f_1 + f_2]t) \tag{2.12}$$

$$\sin(2\pi f_1 t) \cdot \sin(2\pi f_2 t) = \frac{1}{2}\cos(2\pi[f_1 - f_2]t) - \frac{1}{2}\cos(2\pi[f_1 + f_2]t) \tag{2.13}$$

$$\sin(2\pi f_1 t) \cdot \cos(2\pi f_2 t) = \frac{1}{2}\sin(2\pi[f_1 - f_2]t) + \frac{1}{2}\sin(2\pi[f_1 + f_2]t) \tag{2.14}$$

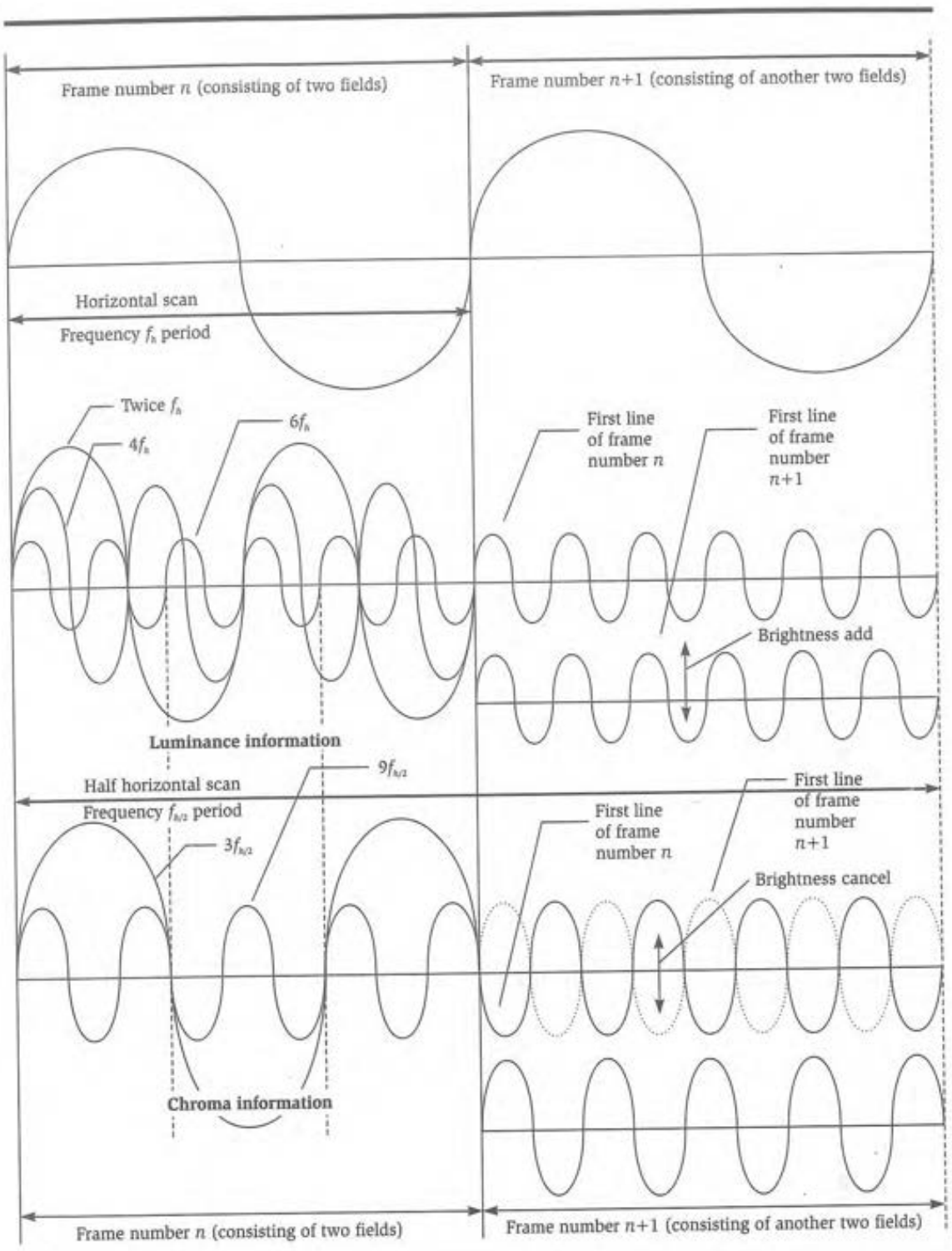


Figure 2.17 Time domain view of the frequency interleaving principle used to hide color information.

Let $f_1 = f_2 = f_c$, which is the carrier frequency for the color information. Then, from Equation (2.12),

$$\cos^2(2\pi f_c t) = \frac{1}{2}\{1 + \cos(2\pi[2f_c]t)\} \tag{2.15}$$

And from Equation (2.13),

$$\sin^2(2\pi f_c t) = \frac{1}{2}\{1 - \cos(2\pi[2f_c]t)\} \tag{2.16}$$

And from Equation (2.14),

$$\sin(2\pi f_c t) \cdot \cos(2\pi f_c t) = \frac{1}{2}\sin(2\pi[2f_c]t) \tag{2.17}$$

If we take the color information signal, $c(t)$, of Equation (2.11) and multiply it by $\cos(2\pi f_c t)$, a cosine at the color information carrier frequency, we get

$$c(t) \cdot \cos(2\pi f_c t) = I(t) \cdot \cos^2(2\pi f_c t) + Q(t) \cdot \sin(2\pi f_c t) \cdot \cos(2\pi f_c t) \tag{2.18}$$

Applying Equation (2.15) to the $I(t)$ term and Equation (2.17) to the $Q(t)$ term, we get

$$c(t) \cdot \cos(2\pi f_c t) = \frac{1}{2}I(t) \cdot \{1 + \cos(2\pi[2f_c]t)\} + \frac{1}{2}Q(t) \cdot \sin(2\pi[2f_c]t) \tag{2.19}$$

$$c(t) \cdot \cos(2\pi f_c t) = \frac{1}{2}I(t) + \frac{1}{2}\{I(t) \cdot \cos(2\pi[2f_c]t) + Q(t) \cdot \sin(2\pi[2f_c]t)\} \tag{2.20}$$

The first term on the right-hand side of Equation (2.20) is the $I(t)$ signal at baseband frequencies. The bracketed terms are modulated onto carriers at twice the chroma carrier frequency, f_c, and can easily be removed with a low-pass filter. Thus, the $I(t)$ signal has been completely separated from the $Q(t)$ signal through quadrature detection.

If we take the color information signal, $c(t)$, of Equation (2.11) and multiply it by $\sin(2\pi f_c t)$, a sine at the color information carrier frequency, we get

$$c(t) \cdot \sin(2\pi f_c t) = I(t) \cdot \sin(2\pi f_c t) \cdot \cos(2\pi f_c t) + Q(t) \cdot \sin^2(2\pi f_c t) \tag{2.21}$$

Applying Equation (2.17) to the $I(t)$ term and Equation (2.16) to the $Q(t)$ term, we get

$$c(t) \cdot \sin(2\pi f_c t) = \frac{1}{2}I(t) \cdot \sin(2\pi[2f_c]t) + \frac{1}{2}Q(t)\{1 - \cos(2\pi[2f_c]t)\} \tag{2.22}$$

$$c(t) \cdot \sin(2\pi f_c t) = \frac{1}{2}Q(t) + \frac{1}{2}\{I(t) \cdot \sin(2\pi[2f_c]t) - Q(t) \cdot \cos(2\pi[2f_c]t)\} \tag{2.23}$$

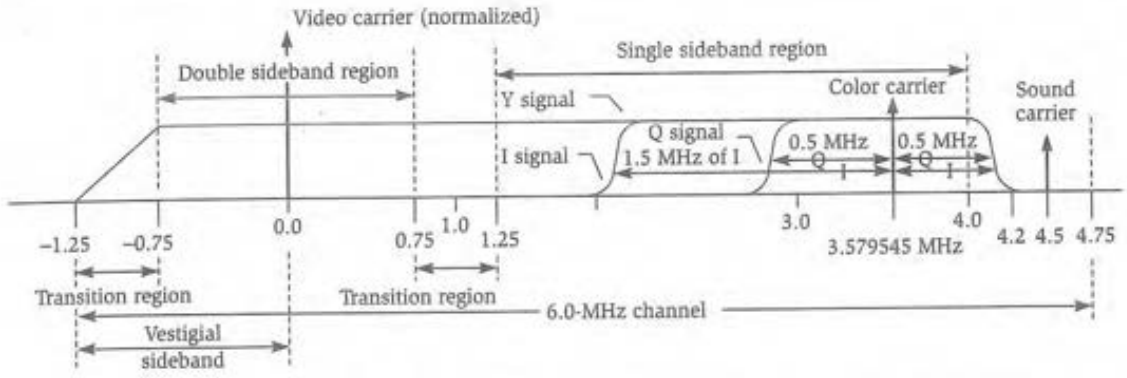


Figure 2.18 Location and bandwidths of the luminance and chroma signals.

As with the $I(t)$ signal, the $Q(t)$ signal has been completely separated through quadrature detection.

The I signal is the *in-phase* color signal and is amplitude modulated onto a color subcarrier that is located at 3.579545 MHz, the 455th odd multiple of half the line-scanning frequency. The Q signal is the *quadrature* color signal. It is amplitude modulated on a carrier that is derived by a 90° phase shifting of the color subcarrier (see Figure 2.18). Note that the Q signal is a double sideband signal, whereas the I signal has a 0.5-MHz double sideband component and an additional 1.0 MHz of single sideband modulation. However, as already mentioned, practical consumer receivers limit bandwidth in both channels to around 0.5 MHz.

It will be recalled from the study of geometry that there are two common forms of coordinate systems, rectangular and polar. In the rectangular system, a location is defined by the distance in the horizontal direction and the distance in the vertical direction. In the polar system, a location is defined by a distance along an angle. This is like giving driving directions of 3 miles east and 4 miles north. If the traveler has an airplane, he can simply be told to fly 5 miles on the heading of 30°. The diagram of Figure 2.19 applies these principles to the quadrature color equation (2.11). The color signal $c(t)$ can be given either as the sum of the $I(t)$ signal and the $Q(t)$ signal or as the resulting magnitude $c(t)$, which is the square root of the sum of the squares of $I(t)$ and $Q(t)$ at an angle given by the inverse tangent of the ratio of $Q(t)$ to $I(t)$. In this "polar" way of interpreting the color signal, the color, or hue, is defined by the angle, and the intensity of the color (termed its *saturation*) is described by the length $c(t)$. Figure 2.20 shows the location of various colors on a polar graph.

The changes made to accommodate color required the horizontal and vertical scan rates to be slightly modified. This is because the difference between the color subcarrier frequency and the aural subcarrier frequency is a frequency well within the video bandwidth. Any nonlinearity in that signal path will produce a

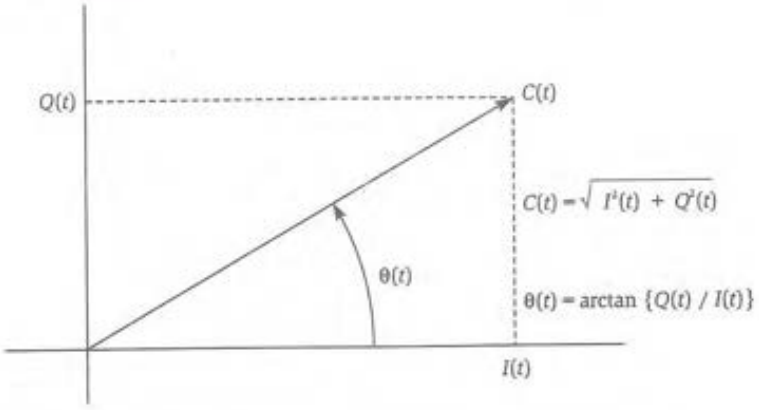


Figure 2.19 Rectangular versus polar representation.

	C	θ
(B–Y)	1	0°
Q	1	33°
Magenta	82	61°
(R–Y)	1	90°
Red	88	103°
Yellow	62	167°
Burst	1	180°
Green	82	241°
Cyan	88	283°
–I	1	303°
Blue	62	347°

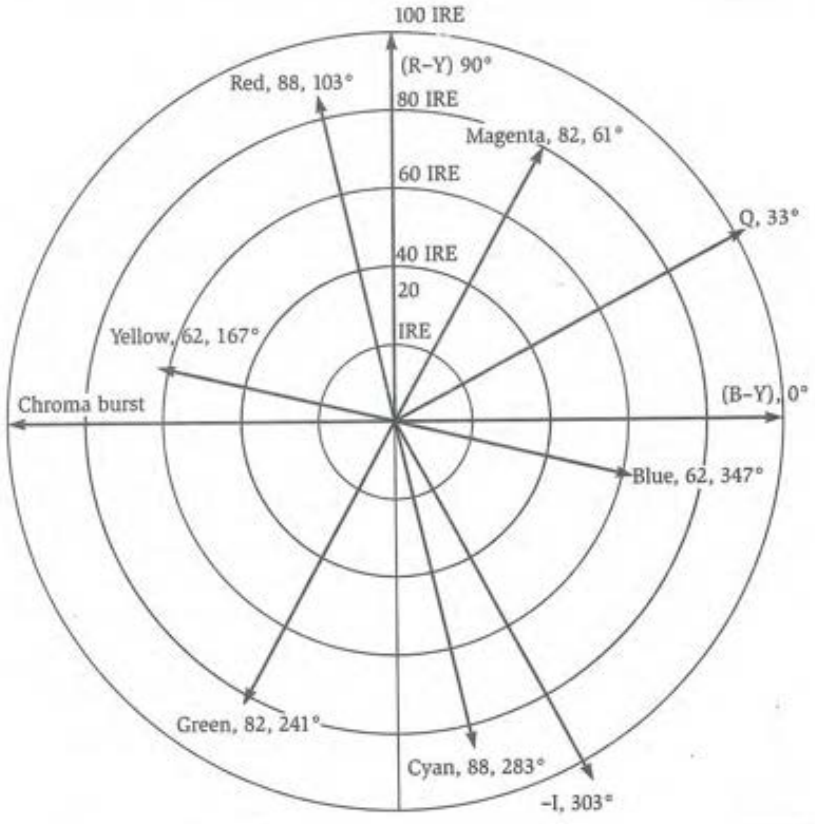


Figure 2.20 Polar representation of television colors.

beat product that is in the video passband and can produce an annoying pattern on the screen. This difficulty has been termed the 920-beat problem because the difference between the frequencies selected for the chroma subcarrier and the sound subcarrier is 920 kHz. Two techniques are used to minimize this problem. The existing 4.5-MHz trap strongly attenuates the aural carrier frequency before the video path. However, it was thought that this measure may not be sufficient in existing monochrome receivers that were designed and produced before the color standard was established. The remaining 920-kHz beat was made less visible by causing it to be frequency interleaved (just like the chroma signals) with the luminance signal. This required that the 920-kHz beat be made an odd multiple of half the scan frequency. To accomplish this (since the chroma subcarrier is already an odd multiple of half scan line frequency), the sound carrier must be any multiple of the horizontal scan frequency (that is, an *even* multiple of half scan line frequency). The 286th multiple of the monochrome horizontal scan frequency (15.75 kHz) is 4.5045 MHz. Using these numbers would require changing the aural subcarrier by 4.5 kHz. Some of the then-existing monochrome television receivers' audio circuits might have failed to perform with this much of a shift in frequency, jeopardizing the compatibility mandate. Also, the effectiveness of the aural frequency trap would be reduced. The alternative was to change the horizontal frequency so that it was the 286th submultiple of the original 4.5-MHz aural subcarrier. This yields a horizontal frequency of 15,734.26 Hz. Since there are 525 scan lines in a frame and two fields in a frame, the field rate will be the horizontal rate divided by 525 and multiplied by 2, yielding 59.94 Hz. The chroma subcarrier is the 455th multiple of half of this line rate, or 3.579545 MHz.

Synchronous detectors in receivers are used to completely separate the two chroma signals. Then, using circuits that perform the functions listed in Equations (2.24) through (2.26), the luminance signal E_Y, the in-phase color signal E_I, and the quadrature signal E_Q are matrixed to yield the blue signal E_B, the red signal E_R, and the green signal E_G that drive the tricolor picture tube.

$$E_B = E_Y + 1.70E_Q - 1.10E_I \tag{2.24}$$

$$E_R = E_Y + 0.62E_Q - 0.96E_I \tag{2.25}$$

$$E_G = E_Y - 0.65E_Q - 0.28E_I \tag{2.26}$$

Synchronization of the color oscillator used in the synchronous detection process is facilitated by the addition of a burst of unmodulated color subcarrier on each horizontal blanking pulse after the horizontal sync pulse (see Figure 2.21).

These techniques allow color television to be compatible with monochrome television. This compatibility is not complete because monochrome receivers built before the introduction of color had video bandwidths of up to 4.2 MHz, providing very sharp black-and-white pictures. When color signals were introduced, these receivers suffered from "dot crawl." The color signal was not

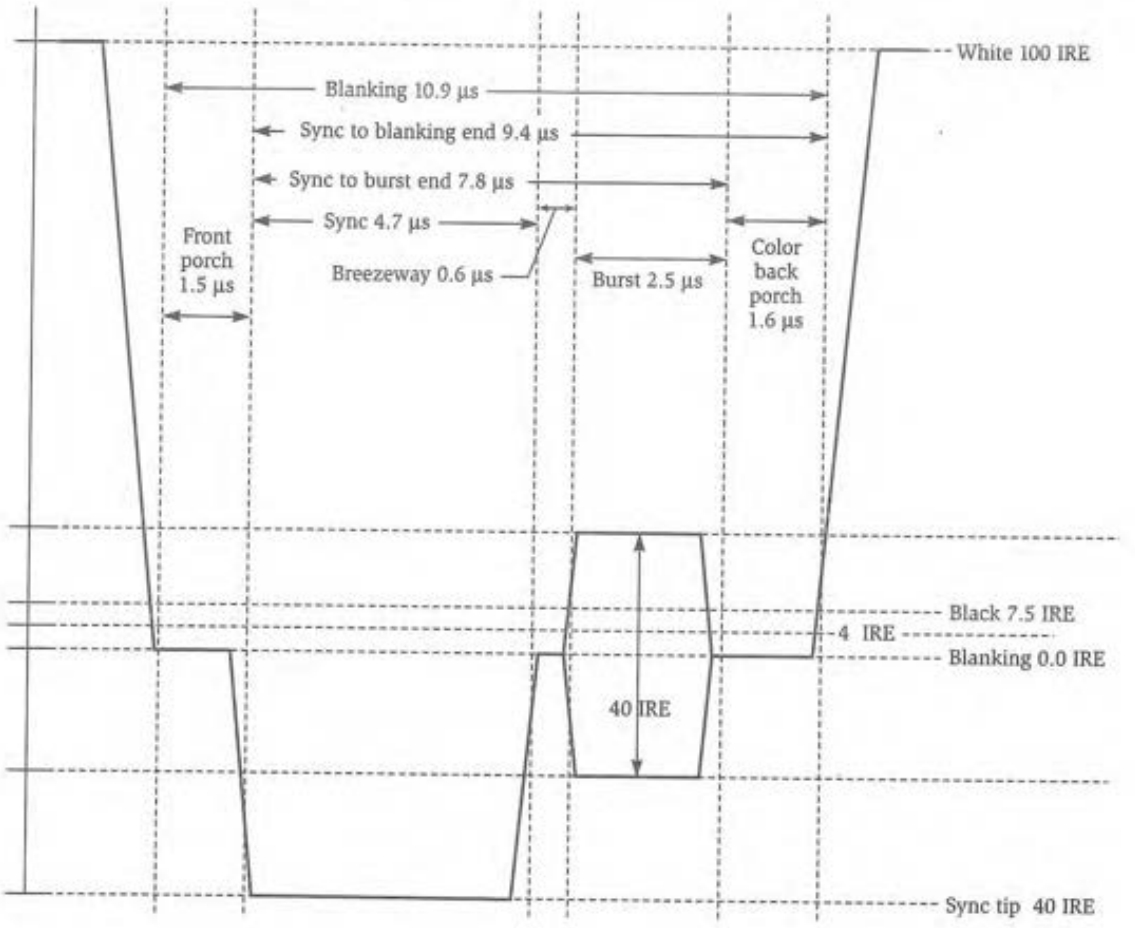


Figure 2.21 Color burst relationship to sync pulse.

adequately rejected by the older receivers and appeared as a moving pattern of annoying dots. This was overcome in later monochrome receivers by introducing a notch in the frequency response that eliminated much of the color signal but reduced sharpness. Alternatively, the video bandwidth of monochrome receivers was rolled off, with much the same effect.

Adequate compatibility was critical to the rational introduction of color television into a market already populated with monochrome television receivers. Consumers who purchased color receivers derived more benefits from the same signals that consumers with monochrome receivers continued to enjoy. Those who could not afford a color receiver could buy a new monochrome receiver and still have access to television. No one was disenfranchised by the technological advance to color.

2.4.2 Stereo Sound

The next improvement to NTSC came in the area of sound.[17] Television sound is frequency modulated on a separate carrier that is a fixed 4.5 MHz above the visual carrier. When stereo sound was added to the television system, the requirement of compatibility was again enforced. This avoided the chaos that would have resulted from making obsolete the existing sound system. Just as with compatible color, there were compromises to monaural receivers when stereo sound was added. But the net benefit to consumers was considered positive. The marketplace gave its approval to both compatible color and compatible stereo sound.

Stereo sound is implemented by first creating a spectrum that includes the sum of the left and right sound channels at baseband. As in FM radio broadcasting and monophonic television practice, the sum signal is preemphasized using a 75-microsecond preemphasis curve. (See the discussion in Chapter 8 under modulators.) The difference between the left and right channels is double sideband suppressed carrier modulated onto a carrier at twice the horizontal scan frequency ($2 \times 15{,}734 = 31{,}468$ Hz). In order to improve the signal-to-noise ratio, this difference signal is *companded*; that is, the signal is compressed in amplitude at the transmission point and expanded at the receiver. The compression is frequency selective and is based on an algorithm originally proposed by the dBx Corporation, whose name it bears. This is in contrast to the FM radio stereo system, which is not companded but employs the same deviation of the sound carrier, by both the sum and difference carriers. Both changes to the television system were brought about by the need for improved signal-to-noise ratio. A limited-bandwidth (10-kHz) monaural second audio program (SAP) channel is frequency modulated onto a carrier at five times the horizontal scan frequency. The SAP channel is intended for a second language or other such purposes. A very narrow-bandwidth (3.4-kHz) "professional channel" is frequency modulated onto a carrier at six and one-half times the horizontal scan frequency. It is used for television plant intercommunications. This entire complex spectrum is then frequency modulated onto the 4.5-MHz carrier (see Figure 2.22). The relationship between the video and aural carriers is tightly controlled, since nearly all television receivers depend on this relationship. The visual carrier is used as the local oscillator to bring the aural spectrum down to baseband. This technique is called the *intercarrier-sound method* of TV receiver design. Since the final aural modulation process is that of frequency modulation, the TV receiver uses a "limiter" circuit to strip off any amplitude modulation that may have become impressed upon the aural signal. Typical sources of this amplitude modulation are noise, particularly impulse noise, and cross modulation in nonlinear elements in the signal path. The TV receiver's aural system then becomes relatively insensitive to amplitude modulation on the aural carrier.

A variation of the previously described aural recovery technique is the *direct-sound method*. It has also been called the split-sound method. In this approach, the aural component is extracted from the composite IF signal prior

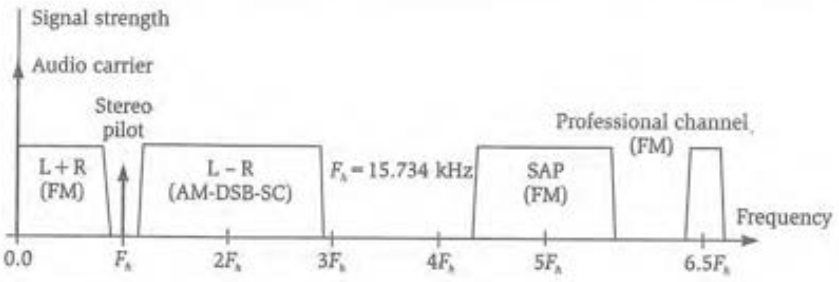


Figure 2.22 Stereo audio baseband spectrum.

to video detection. This signal is subjected to limiting to remove incidental amplitude modulation components and presented to an FM detector.

When direct-sound techniques are applied, the aural signal is not dependent upon any properties of the visual carrier, as it is in the intercarrier-sound approach, but is now affected by the accumulated instability of processing oscillators (heterodyning) along the signal chain. These sources of instability include the transmitter, cable heterodyne processors, and local oscillators of intervening frequency translation devices, such as cable subscriber terminal devices and the tuner of the TV itself.

There are trade-offs in making a choice between intercarrier-sound and direct-sound designs. Some visual modulators suffer from nontrivial amounts of incidental phase modulation of the visual carrier. The intercarrier principle causes these visual carrier instabilities to be transferred to the aural carrier, and this can introduce sound distortions. A direct-sound design would not be subject to this problem. Essentially all consumer TVs and VCRs use the intercarrier principle.

2.4.3 Ancillary Signals

It was determined that more uses can be made of the electron blanking time. One important application is to use the VBI to carry analog test signals. The analog test signals can be used to measure the transmission characteristics from the signal source to intermediate points along its path to the final point of use. Another use for the VBI is to carry analog signals that represent digital data. The data signals can be of two or more levels that are resolved into data bits by appropriate circuits. Since the "digital" signals are of just a few discrete levels, the data detection circuits can discriminate against significant amounts of noise, distortion, and interference. This makes these data signals more robust than the analog video signal itself for most forms of interference.

The first U.S. attempt to use the VBI for ancillary data purposes was in 1970, when the National Bureau of Standards (NBS—now the National Institute of

Standards and Technology, NIST) proposed to use it for the distribution of precise time information nationwide. The American Broadcasting Company (ABC) television network was a partner in that effort. Even though this initiative did not result in service, ABC recommended a captioning service for the hearing impaired.

Eventually the FCC reserved line 21 of field one of the television signal for the transmission of *closed captioning* (CC) in the United States in 1976. In 1990, Congress passed the Television Decoder Circuitry Act, mandating that after July 1, 1993, new television receivers of 13-inch diagonal display measure or greater must include caption decoding circuits. Approximately 20 million television receivers per year are covered by this requirement. In 1992, NCI worked with the FCC and the Electronic Industries Association (EIA) to develop captioning technical standards. The 1996 Telecommunications Act required the FCC to institute rules requiring closed captioning on video programming but allowing exemptions for programming that would suffer an "undue burden."

The CC system supplies data to appropriate digital and analog circuits that place carefully timed text on the television screen, making it possible for the hearing impaired to read a description of the conversation taking place and have indications of other relevant sounds. Similarly, those who cannot understand the spoken words may have text in their language, allowing them to follow the program. The CC system uses very low-speed data in order to minimize the impact of transmission path problems such as reflections and interfering signals. The data rate for the CC system is 503.5 kb/s of binary (two-level) data. This allows only two 8-bit characters to be transmitted per VBI line. If only field one is used, there will be 30 of these lines per second. This yields 480 bps, or 3,600 characters per minute. If the average word is 5 characters long with a space, then 600 words can be conveyed per minute. The rest of the line is occupied with both a burst of seven sine wave cycles of 503.5-kHz clock run-in and a unique "start bits" pattern placed at the beginning of the line. These signals synchronize the detector circuitry. Since only line 21 is protected for captioning by FCC rule, the rate of transmission is slow, but it is perfectly adequate for the purpose. The on-screen display consists of a maximum of 15 rows of 32 characters each. The captions usually appear only on rows 1 through 4 and rows 12 through 15. The middle rows are usually transparent to show the action. A text mode provides scrolling text. Further details can be found as part of EIA standard number EIA-608.

The closed captioning signal carries four components. There are two captioning channels and two text channels. The first captioning channel is synchronized to the video programming, causing the words to carefully match the video. The second captioning channel is not synchronized.

The FCC expanded the captioning standard EIA-608 to allow use of line 21 of field two. This adds two more captioning channels and two more text channels. A fifth channel has been added to carry extended data services (EDS). EDS can carry a wide variety of additional information, including precise time information to set clocks in consumer products. Finally, the original intent of the NBS will be realized.

The channel's name and call letters are included along with current program information, such as title, length, rating, elapsed time, types of audio services and captioning services, and intended aspect ratio. Also included is the data for the "V-chip" (violent programming advisory), which is intended to control children's access to programming that parents deem objectionable. Public service announcements such as weather and emergency advisories will be transmitted. Cable system channel layout information will be provided, causing the channel number indicator to use the more familiar channel identification number rather than the number associated with the frequency utilized. This facility can bring the same "channel mapping" benefits subscribers have enjoyed in their cable set-top terminals to consumer electronics products.

A subsequent VBI data transmission system, called *teletext*,[18] was invented to provide ancillary services to television users. The teletext system can display up to 24 rows of 40 characters on the television screen (but a specification of 20 rows was selected for use in the United States). Teletext quickly evolved into a transmission system for more complex data, including the downloading of software to computers. At the time teletext was introduced, electronic components were still relatively expensive, which severely limited the affordable capability of the teletext systems.

Teletext is a more aggressive form of data transmission than CC. It has been successful in Europe but has failed to enjoy commercialization in the United States. Teletext originated in Great Britain, with experimental transmission commencing in 1972. The British Broadcasting Corporation (BBC) branded their service Ceefax, and the Independent Broadcast Authority (IBA) called their service Oracle. France developed a packet-based teletext system called Antiope, based on a transmission system called Didon. Later, Canada developed another system, called Telidon, which featured higher-resolution graphics. The Japanese system, called Captain, featured "photographic coding" to accommodate Chinese Kanji characters and the Japanese Kana set.

There are a number of reasons for the difficulties in the United States. Principal among these was the failure to find a strategy that made money. Without this, the system could not be supported. Additional difficulties included the high cost of memory at the time of implementation. Even though a teletext page requires only about a kilobyte of storage, that small amount of memory was expensive then. Further problems centered on the quality of the graphics. The less expensive World System Teletext (WST), based on the British approach, had crude "Lego-style" graphics in its basic form. The other contender, the North America Presentation Layer Protocol System (NAPLPS), used a higher-resolution graphics system that painted itself on the screen painfully slowly, trying the patience of the user. Still another complication was the FCC's 1983 decision to allow two standards and hope that the marketplace would resolve the confusion. One of the systems is WST, and the other is the NAPLPS evolution of Antiope, Telidon, and efforts by AT&T. A final problem was reliability of data reception. In a test in the San Francisco Bay Area, only about 25% of installations of the

NAPLPS system were trouble free. The remainder suffered from various degrees of multipath impairment. The more robust WST system was not tested in that environment.

Both U.S. teletext systems have a data rate of 5.727272 Mbps, which is 364 times the horizontal rate and 8/5 of the color subcarrier. The data signal has a nonreturn to zero (NRZ) binary format. The WST data line consists of eight cycles of clock run-in (16 bits) followed by a unique 8-bit "framing code" followed by 16 bits of control codes and a payload of 32 8-bit display words. The displayed row has 40 characters, 8 more than can be conveyed in one VBI line of data. This is accommodated by transmitting 4 rows of 32 characters on 4 VBI lines and a supplementary VBI line with the 4 groups of 8 characters needed to finish the previous 4 display rows. This process is called *gearing*, because of the ratios between the display and the transmitted VBI data. Since the page format is 40 characters by 20 rows, with an additional "header row," 21 field lines are required plus 5 field lines of supplementary 8-character groups, for a total of 26 field lines per full page of WST teletext. The payload of 296 bits per line allocated means that if one VBI line in each field is allocated, a data rate of $296 \times 2 \times 30 = 17{,}760$ bps is obtained. Ten lines of VBI are possible (line 21 is reserved for captions, and the first nine lines form the vertical synchronization pulses), yielding a maximum of 177.6 kb/s for full VBI utilization.

The WST system maps the data location in the VBI line to memory locations and to screen locations, and always puts data in the same memory place. This allows for a very simple error-protection scheme. Since the instructions in the header are Hamming-code protected, a measure of the quality of the received signal is obtained. If the signal is of low quality, it is not loaded into memory. Only good-quality data is stored. As a result, good data can be accumulated from repetitions of the page until a good page of data is built up. It is also possible to use a "voting" approach to obtain very robust transmission.

The fundamental difference between the WST and the evolving set of Antiope, Telidon, and NAPLPS systems is that the latter systems all used a packet structure. They have been characterized as asynchronous because there is no mapping between the transmission scheme and memory and screen locations.

The Public Broadcasting System (PBS) has developed a packetized data delivery system based on teletext called the PBS National Datacast Network. The standard teletext data rate of 5.72 Mbps is used, yielding 9,600 baud per VBI line allocated per field. There are a wide variety of commercial applications for this signal. Currently, the StarSight Electronic Program Guide (EPG) signal is distributed via PBS.

A system by the WavePhore Corporation utilizes a teletext-like system in lines 10 through 20 in each field, for a data speed of up to 150 kb/s. WavePhore added substantial error-detection and -protection bits to its structure to protect against multipath and other transmission problems.

Underutilized portions of the NTSC spectrum can be employed to "hide" data. In many cases, the process of hiding the data is incomplete, and results in

artifacts under certain conditions. In other cases, the preparation of the NTSC signal to better hide the data reduces video quality. So the challenge is to both hide the data and not impair video quality while retaining signal robustness and the potential for an economic implementation.

The National Data Broadcasting Committee (NDBC) was formed in 1993 to establish a single standard for data transmission in video. On June 28, 1996, the FCC approved digital data transmission in the video portion of broadcast television transmission in its Report & Order (R&O) "Digital Data Transmission Within the Video Portion of Television Broadcast Station Transmissions," in MM docket No. 95-42. It allows four formats: two of the formats, by Yes! Entertainment Corporation and A. C. Nielsen Co., place low-data-rate signals in the overscan region of the picture; the other two systems, Digideck and WavePhore, embed the digital signal into the video signal.

The WavePhore system begins by reducing video luminance and chrominance bandwidths. The "luminance" is reduced from its theoretical value of 4.2 MHz to a lower value of 3.9 MHz, and the upper sideband of the color signal is reduced by about 300 kHz. It is then possible to insert a data signal in this region at a carrier frequency of approximately 4.197 MHz above the video carrier and a strength approximately 20 dB above the noise floor of the video system. The data is synchronous with the video carrier and thus with the horizontal line frequency. As an odd multiple of one-quarter of the horizontal scan frequency, it interleaves between the luminance and chrominance bundles of spectral energy. Data is not sent during the vertical and horizontal blanking intervals. Thirty bits of data are sent per video line. There are 240 available lines per field (not counting the VBI, during which the signal is blanked). This yields a raw data rate of 431.6 kb/s. After error-correction coding, the raw data rate is reduced to approximately one-fourth of a telephone T1 rate, or 384 kb/s, and referred to by WavePhore as their system TVT1/4.

WavePhore shuffles the data before applying biphase modulation and filtering out the lower sideband. Shuffling the data reduces its visibility in the video. An adaptive equalizer is used in the receiver. A major advantage of the WavePhore approach is that once inserted into the video, the data can be conveyed through the video path without giving it further attention. The WavePhore VBI system and the WavePhore subvideo system can be combined to provide more than 500 kb/s.

There is some degradation of pictures using this system. However, the FCC is willing to let the broadcaster determine what its individual marketplace values are and to respond to that decision.

The Digideck system adds a differential quadrature phase shift keyed (DQPSK) signal carrying about 500 kb/s placed 1 MHz below the video carrier. In this regard, it is similar to the European NICAM system for adding digital audio to analog television broadcasts (see Chapter 7). This places the new carrier in the vestigial side band region of the signal. To accommodate this, the lower VSB slope is increased. Rather than starting at the traditional 750 kHz below

picture carrier, in the Digideck system it starts at 500 kHz and drops more rapidly. The carrier is about 36 dB below peak power and has a raw capacity of 700 kb/s. Forward error correction and other overhead burdens reduce the data capacity to around 500 kb/s. Digideck calls the new carrier the D-Channel. The data signal is clocked synchronously to the television signal for ease of recovery and for better hiding in the video. The Digideck receiver also depends on an adaptive equalizer. A consequence of the D-Channel is that it must be inserted at the transmitter site and be brought there by an alternative path. Like the WavePhore system, Digideck introduces some artifacts. A marketplace approach will allow the broadcaster to determine acceptability.

The Yes! Entertainment Corporation's system introduces a pulse in the video between 9.1 and 10.36 microseconds after the start of the horizontal synchronization pulse. The data rate is very low, about 14 kb/s. Its application is to deliver audio to a talking toy teddy bear! A. C. Nielsen uses line 22 of one field of the video for transmitting a program source identification. This ID is used to measure the viewing population for statistical purposes.

EnCamera Sciences Corporation has presented papers on a system that compatibly embeds up to 4.5 Mb/s in an analog 6-MHz NTSC television signal. Up to 3.0 Mb/s is modulated using multilevel signaling onto a carrier that is in quadrature with the visual carrier. This approach would normally result in visible artifacts in the analog video. However, techniques have been developed that avoid these problems. An additional 1.5 Mb/s is amplitude modulated onto the aural carrier using multilevel signaling and specialized filters.

2.5 Signal Quality and Customer Expectations

The human response to television pictures is a complex phenomenon. This makes establishing performance specifications difficult. Yet these specifications are important for setting design goals, purchasing objectives, contractual obligations, and regulations.

Psychophysical testing assesses subjective responses of a large number of subjects to television pictures and assigns figures of merit versus objectively measured parameters such as random noise, phase noise, microreflections, composite triple beat, and chroma/luma group delay.

2.5.1 CableLabs Updates Early Testing

Early testing was done by the Television Allocations Study Organization (TASO) in 1958.[19] Further testing was done in 1983 by the network labs of the Columbia Broadcasting System (CBS), the National Aeronautics and Space Administration (NASA), and the Canadian Research Council (CRC).[20] Most recently, in 1991, work sponsored by CableLabs added to the base of knowledge concerning signal quality and customer expectations.[21]

The most significant conclusion is that customer expectations are rising. The TASO study proclaimed that in 1958 consumers found a weighted signal-to-noise ratio of about 30 dB to be "somewhat objectionable." In 1983, subjects' expectations were for a much better picture. About 40-dB signal-to-noise ratio was considered to be "somewhat objectionable." The CableLabs study found 45–46 dB to be considered "slightly annoying." This is a very substantial increase in demand for picture quality.

The CableLabs tests were conducted under carefully controlled circumstances and used both expert and nonexpert viewers. The expert viewers served primarily to set the endpoints of the scales for the different parameters. The selected visual material was presented on a custom-made analog video disc and on D2 tape.

Well-established procedures defined by the International Consultative Committee for Radio (CCIR) were employed.[22,23] The subjective scale ranged from "imperceptible," to "perceptible, but not annoying," "slightly annoying," "annoying," and "very annoying." The double-stimulus pair-comparison method used allows subjects to compare pictures to reference images. This eliminates the need to remember what images looked like at other points in time.

Subject material was selected for sensitivity to the subjective impairments under study. Sources for the material included the Society of Motion Picture and Television Engineers (SMPTE), the European Broadcast Union (EBU), and Kodak Corporation. All these entities have long histories in the analysis of images.

Five impairments were studied in the CableLabs test: (1) random noise, (2) phase noise, (3) chroma/luma delay, (4) microreflections, and (5) composite triple beat (CTB). These impairments are briefly described here (see Table 2.1). Random noise is the result of thermal motion of electrical charges and yields "snow" in the picture. Phase noise is the visual manifestation of instability of oscillators that are used in heterodyning devices to shift frequencies. Several of these systems are in the signal path from point of origination to the subscriber's screen. Phase noise usually appears as near-horizontal streaking of the picture. Chroma/luma delay results in the colored part of the image misregistering from the luminance part of the picture. The color is smeared and located in the wrong place, giving rise to a comparison with comic book drawings. Microreflections

Table 2.1 Suppression of Impairments for Various Subjective Reactions

Impairment	*Imperceptible*	*Perceptible, but not annoying*	*Slightly annoying*	*Annoying*
Random noise	53 dB	48–51 dB	45 dB	40 dB
Phase noise		90 dB	85 dB	
CTB		50 dB	40 dB	30 dB
Chroma/luma delay			300 ns	400 ns

cause delayed versions of the signal to add destructively or constructively. In many cases, the picture is blurred. In a few circumstances, the subjective sharpness of the picture may appear to improve even though the objective measurements may lead to a different conclusion. Composite triple beat (CTB) is the creation of new, undesirable frequency components due to the signal's passing through nonlinearities. The nonlinearities cause new frequencies to be created that interfere with the quality of the picture.

Table 2.1 summarizes the results. Quantitative data was not taken for microreflections and so is not included in the table. However, microreflections below 20 dB were not of consequence to most viewers. Below 10 dB, some pictures looked sharper, whereas others looked softer. The most important conclusion is that subscribers have become more critical and are likely to become even more so as they see more images delivered with digital television (DTV).

2.5.2 The W Curve and Direct Pickup (DPU)

In 1986, CBS Laboratories and the NAB tested subjective reaction of viewers to co-channel interference.[24] The desired signal was at broadcast channel 23 (visual carrier at 525.5 MHz) and consisted of three images: a gray flat field, a still image (of a boat), and moving video. The undesired signal was a series of sine waves at intervals of 500 kHz, starting at the frequency of the lower band edge of the channel. The ratio of the strength of the interfering sine wave to the strength of the visual carrier was adjusted until expert viewers agreed that the interference was "just perceptible." To be "just perceptible" at the lower band edge, the strength of the interfering sine wave needed to be attenuated so that it was only some 15 dB below the visual carrier strength. As the frequency of the interfering carrier was increased, more attenuation was required to bring it down to levels that were "just perceptible." Maximum attenuation, some 68 dB, below visual carrier strength was required at a frequency 250 kHz above the visual carrier. Less attenuation, about 55 dB, was required at the middle of the channel. Again more attenuation, in the low 60s, at the chroma carrier was required. And then, as the upper band edge was approached, less attenuation was required. These measurements yielded a W-shaped curve, as shown in Figure 2.23. The video content of the desired signal had some influence on the amount of attenuation required in the middle of the channel. Somewhat surprisingly, the flat field hid the interference better in the low to mid-frequencies, whereas the moving picture hid the interference better around the color subcarrier. Most surprising was the finding that the difference between the "just perceptible" level and unacceptable levels was just 2–3 dB.

Although co-channel interference is not generally a problem for cable television, direct pickup (DPU) interference is a major headache. In the case of a strong broadcast channel that is carried on the same (or close) frequency on cable, the DPU may result in a ghost or smearing of the image, since the

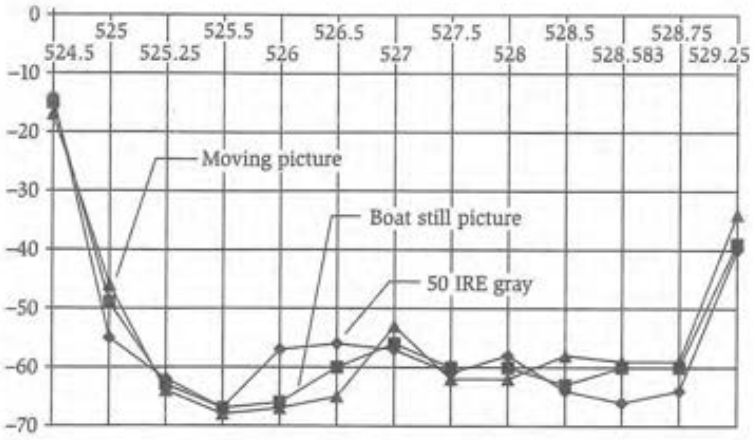


Figure 2.23 W curve: Level of "just perceptible" interfering carriers for three images.

propagation time through the cable is longer than through the air. In the case of other users of the spectrum, second-level-order distortion products, local UHF broadcasters, or cable systems using other than off-air frequencies, interfering patterns may appear on the screen. One manifestation of this is in the cable channel 19 area, where pager transmitters are authorized to operate. Direct pickup can occur in the cable plant or in equipment connected to the cable drop. Those DPU signals can propagate between products connected to the same drop and cause picture degradation. If the tuner of a TV or VCR is not adequately shielded, the DPU will cause ghosts and other interference. An occasional problem arises when subscribers connect FM radios to the cable. These products generally have almost no shielding on their tuners and may conduct DPU signals back into the cable system to interfere with reception by other devices. The W curve gives an indication of sensitivity to DPU at various frequencies.

In 1993, the 1986 CBS tests were repeated, with emphasis on their application to DPU. Since the TASO observations were found to no longer be applicable, it was reasonable to wonder whether consumers had become more sensitive to co-channel and DPU problems. The tests were done on three cable channels (6, 12, and 78), at four frequencies relative to visual carrier (+0.25, +0.75, +1.75, and +2.55 MHz), at two desired signal strengths (0 and +10 dBmV), and at two carrier-to-noise ratios (43 and 50 dB). Only the gray flat field was used for the desired signal. The 1993 tests generally confirmed the results of the 1986 CBS experiments. The W shape of the curve remained. Across the three channels, four frequencies, two desired signal levels, and two carrier-to-noise ratios, the average value of the sine wave strength attenuation relative to the carrier strength to achieve "just perceptible" impairment levels was 56 dB. Since these tests involve human subjects and a wide range of variables, they are difficult to repeat and obtain closely similar results.

2.6 Other Television Systems

The NTSC color system was historically the first to be adopted in the world. Because of the close economic relationships between the North American countries and Japan, all these countries adopted the same color television system, allowing products to be made and sold in these countries interchangeably, with only minor modifications due to differing channel frequencies.

Other countries adopted color television later.[25] As a result, they were able to take advantage of the experience of the NTSC countries and make modifications that better suited their priorities. In some cases, these priorities included the desire to protect domestic manufacturing by choosing standards that were different. Controlling the licensing of the intellectual property (patents) erected market-protecting barriers.

In the NTSC system, there are two distinct types of crosstalk between the luma and the chroma signals. Linear crosstalk is a consequence of luma signals (under some circumstances) having spectral components that lie in the chroma regions. The classic example of this is the spurious rainbow pattern that appears on a checkered or tweed jacket. The pattern produces frequencies that stimulate the chroma circuits. Conversely, the 3.58-MHz chroma signals are not completely eliminated from the luma circuits and appear as small moving dots, especially on sharp edges in highly saturated color areas. Slight nonlinearities cause the luminance and chrominance signals to modulate each other. In particular, the chroma phase and amplitude become a function of the signal's brightness. These phenomena, known as *differential phase* and *differential gain*, cause infidelities in color reproduction. As circuit technology has improved, these problems have become less severe. But in the first years, using vacuum tube circuits, great care and expense was required to keep these problems manageable.

The primary objective of European developments in color television was to minimize the impact of differential phase and differential gain and of other undisciplined phase shifts in the signal path. The principle behind the approach is to shift the phase on one of the chroma signals by 90° on adjacent lines. This approach gives rise to the technique's name: phase alternating at line (PAL) rate. In addition, the chroma reference burst also shifts 90° at the end of each line. The beneficial impact of this approach is that hue shifts on adjacent lines are in opposite directions. Two approaches can be used to take advantage of this. In the less expensive and simpler approach, it is left to the observer's eye to average the hues. This is quite effective for the small phase errors of modern equipment. However, larger phase errors give rise to horizontal lines sometimes called Hanover bars or Venetian blinds. The more expensive alternative is to use delay lines to electrically average the signals before display. While avoiding the artifacts mentioned, vertical chroma resolution is halved and additional expense results.

PAL also requires a relationship between the chroma subcarrier frequency and the line scan rate. However, because the phase of the chroma burst changes 90°

at the end of every line, the relationship must be that the subcarrier frequency is an odd multiple of the sum of one-quarter of the scan line frequency plus one-half the frame frequency. This is in contrast with NTSC's simpler relationship of an odd multiple of one-half of the horizontal scan frequency.

The PAL strategy of alternating phase on adjacent lines and the need to create a relationship with an odd multiple of one-quarter the horizontal scan frequency means that the chroma spectral lines are not centered in the luma spectral gaps. In the time domain, the averaging of brightnesses caused by the chroma signal is even less effective. The saving grace is that the chroma subcarrier frequency is almost a megahertz higher than in NTSC, and therefore the chroma dots are much smaller and less visible.

European television uses the UHF frequencies and thus has more bandwidth available for channelization. Whereas the North American and Japanese video baseband bandwidth is 4.2 MHz, the British have 5.5 MHz and most of Europe has 5.0 MHz. The French and some of the former Soviet bloc countries use 6.0 MHz. Additionally, most of the European countries utilize a 50-Hz power line frequency and thus have chosen a vertical monochrome frame rate of 50 frames per second. The consequence of the wider bandwidth and lower frame rate is higher vertical and horizontal resolution. These systems have 625 scan lines per frame, compared to the NTSC's 525 scan lines. The horizontal scan line rate is half of 50 times 625, or 15,625 Hz. However, the lower frame rate gives rise to flickering in bright areas of the image and to reduced temporal resolution. The PAL chroma subcarrier is at approximately 4.43 MHz (compared with NTSC's 3.58 MHz). A further benefit of the wider bandwidth is that both color signals in PAL have the same bandwidth, 1.3 MHz. The NTSC strategy of two different bandwidths for matrixed colors matching the human response to flesh tones is avoided. Simple color difference signals are used involving red minus luma and blue minus luma.

The French SECAM system is more complex and uses frequency modulation for the two color difference signals. These two signals are alternated on adjacent lines, yielding half vertical color resolution.

2.7 Summary

The fundamentals of analog television are covered in a manner intended to build an understanding around which to better appreciate cable television technology. We presented the basic concepts behind the structure of the signal and its modulation. This allows the requirements on the signal processing implemented in cable to be developed in later chapters. Bandwidth compression is important in analog television. The interlaced scanning system and vestigial sideband modulation are both important bandwidth compression techniques. When color is added to the signal, the careful matching of the signal to the human color visual response saves bandwidth.

We surveyed the consequences of the technical compromises forced by the need to have affordable receivers in the early days of television. We explored the important concept of compatibility, demonstrating that complete compatibility is not necessary. Some modest degree of degradation of the performance of earlier products is acceptable in order to advance the performance of new products. A balance is struck.

The story of NTSC is replete with technical compromise. The goals have always been to match image and sound quality to the human who is expected to pay for the hardware and the services. A good compromise and a good bargain have been attained.

Endnotes

1. Donald G. Fink, ed., *Television Standards and Practice: Selected Papers from the Proceedings of the National Television System Committee and Its Panels*. New York: McGraw-Hill, 1943.
2. David E. Fisher and Marshall Jon Fisher, *Tube, The Invention of Television*. The Sloan Technology Series. Washington, DC.: Counterpoint Publishers, 1996.
3. Ibid.
4. Bernard Grob, *Basic Television*, 3rd ed. New York: McGraw-Hill, 1964.
5. K. Blair Benson and Jerry Whitaker, *Television Engineering Handbook*, rev. ed. New York: McGraw-Hill, 1992.
6. Bernard Grob, op. cit.
7. Mischa Schwartz, *Information Transmission, Modulation, and Noise*, 2nd ed. New York: McGraw-Hill, 1970.
8. K. Blair Benson and Jerry Whitaker, op. cit.
9. B. P. Lathi, *Communication Systems*. New York: Wiley, 1968.
10. Mischa Schwartz, op. cit.
11. Donald G. Fink, *Principles of Television Engineering*. New York: McGraw-Hill, 1940.
12. *Proceedings of the IRE, Color Television Issue*, Vol. 39, No. 10, October 1951. The Institute of Radio Engineers, New York.
13. *Proceedings of the IRE, Second Color Television Issue*, Vol. 42, No. 1, January 1954. The Institute of Radio Engineers, New York.
14. *Color TV Training Manual*, 3rd ed. Indianapolis, IN: Howard W. Sams, 1970.
15. A. M. Morrell, H. B. Law, E. G. Ramberg, and E. W. Herold, *Color Television Picture Tubes*. New York: Academic Press, 1974.
16. Keith Jack, *Video Demystified*. San Diego, CA: HighText Publications, 1996.

17. Theodore S. Rzeszewski, *Television Technology Today*. Piscataway, NJ: IEEE Press, 1985.
18. Special Issue on Consumer Text Display Systems, *IEEE Transactions on Consumer Electronics*, Vol. CE-25, No. 3, July 1979.
19. C. E. Dean, Measurements of the Effects of Interference in Television Reception. *Proceedings of the IRE*, Institute of Radio Engineers, New York, June 1960.
20. G. Chouinard, W. A. Whyte Jr., A. A. Goldberg, and B. L. Jones, Experimental Results Supporting the Determination of Service Quality Objectives for DBS Systems. *IEEE Journal on Selected Areas in Communications*, Vol. SAC-3, January 1985.
21. B. L. Jones and J. A. Turner, Subjective Assessment of Cable Impairments on Television Picture Quality, *1991 NCTA Technical Papers*, p. 92. Washington, DC: NCTA.
22. Method for the Subjective Assessment of the Quality of Television Pictures, Recommendation 500-4, *Recommendations and Reports of the CCIR*, Vol. XI, Part 1. Dubrovnik, Croatia: CCIR, 1986.
23. B. L. Jones and L. E. Marks, Picture Quality Assessment: A Comparison of Ratio and Ordinal Scales. *SMPTE J*, Vol. 94, No. 12, December 1985.
24. G. Chouinard et al., op. cit.
25. Geoffrey H. Hutson, *Colour Television Theory, PAL-System Principles and Receiver Circuitry*. London: McGraw-Hill, 1971.

Other References of Interest

Donald Christiansen, *Electronics Engineers' Handbook*, 4th ed. New York: McGraw-Hill, 1996.

Margaret Craig, *Television Measurements, NTSC Systems*. Beaverton, OR: Textronix, 1990.

Jeffrey L. Thomas, *Cable Television Proof-of-Performance*. Englewood Cliffs, NJ: Prentice Hall, 1995.

Ted Rzeszewski, *Color Television*. Piscataway, NJ: IEEE Press, 1983.

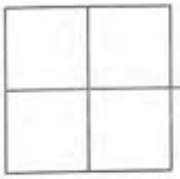
Joel Brinkley, *Defining Vision, The Battle for the Future of Television*. New York: Harcourt Brace, 1997.

J. Allnatt, *Transmitted Picture Assessment*. New York: Wiley, 1983.

B. L. Jones, Subjective Assessment of Cable Impairments on Television Picture Quality, *1992 NCTA Technical Papers*, pp. 1–25. Washington, DC: NCTA.

NCTA Technical Papers, yearly, NCTA, Science and Technology Department, Washington, DC.

IEEE Transactions on Consumer Electronics, quarterly, Institute of Electrical and Electronic Engineers, Piscataway, NJ.



Chapter 3
Digitally Compressed Television

3.1 Introduction

This chapter introduces digital compression, transport, and transmission for television signals, particularly as they relate to cable systems. However, it is worth taking a moment to consider why the cable industry has transformed itself to embrace digital technology. After all, in the mid-1990s the cable industry was starting a hybrid fiber-coax (HFC) upgrade program to expand channel capacity to greater than 100 analog channels. Why then was digital technology adopted so rapidly in the late 1990s after many HFC upgrades had already been completed?

Digital compression was first applied to moving images in video conferencing applications.[1] From there, interest in its application to television signals started with early satellite systems (such as SkyPix and PrimeStar). However, the essential push may have been provided by high-definition television (HDTV).[2] An HDTV signal contains about six times as much information as a standard-definition television (SDTV) signal and consumes at least 12 MHz of spectrum in analog format. This fact makes digital compression a necessity for HDTV. Once the investment had been made to digitally compress an HDTV signal, the same technology was naturally applied to SDTV. The first application was in direct broadcast satellite (DBS) systems — digital compression was a perfect fit for a new system that was constrained by expensive, and limited, satellite transponder capacity. Digital transmission also reduced the required signal-to-noise ratio, allowing for a smaller satellite dish (on the side of the home).

With digital compression, DBS systems could offer enough channels to pose a competitive threat to cable systems, especially those that had not done an HFC upgrade. The range of programming continued to expand, and suddenly even 100 analog channels wasn't enough to carry them all. Moreover, the DBS industry began to actively market its "digital picture and sound quality." As a result, the cable industry turned its attention to digital. The new digital tier allowed cable to lift its per-subscriber revenue by adding new premium subscription content, and the pay-multiplex was created by HBO. In addition, digital pay-per-view greatly expanded choice for the customer. Finally, digital compression allowed an upgraded system to offer video-on-demand, something DBS could not offer.

Traps and analog scrambling are relatively easy to circumvent, and signal theft has become a significant issue in analog systems. Digital technology enables a more secure conditional access mechanism based on *cryptography* (see Chapter 21).

This chapter will focus on the methods employed by the satellite, cable, and broadcast systems, all of which use MPEG-2 video compression. In addition, they are all based on a one-way, time division multiplexed approach known as *MPEG-2 transport*, although there is growing interest in alternative methods that are more suitable to Internet Protocol (IP) networks under the umbrella term *streaming media*.

We will consider video compression and then audio compression before delving into multiplexing and data transport of compressed digital content over existing satellite, cable, and broadcast systems. First, what is the history of broadcast digital television?

3.2 Broadcast Digital Television

The NTSC television system was created at a time when electronics were expensive and signal processing difficult and limited. The programs were few, and they were relatively expensive to produce. An audience was lacking. A great many compromises were made to allow cost-effective receivers. Without affordable receivers, there would be no point in having efficient use of the spectrum. The priority was rapid penetration of the marketplace and the generation of demand for programming and receivers.

The situation is now quite the opposite. The scarce resource is spectrum. Electronics and electronic processing are affordable and on a predictable path of further cost reduction (according to Moore's law; see Section 3.2.2). The goal now is to create systems that efficiently utilize the spectrum and that will last for several decades without becoming painfully obsolete. The first step is to study the redundancies in television images that can be eliminated to lighten the spectrum burden.

Interest in digital television goes back several decades. Long before digital television was practical for consumer use, it held interest for military and professional applications. Digital signals have some important advantages:

- The ability to transmit signals over arbitrarily long distances (using repeaters if necessary; see section 3.2.2)
- The ability to store the signals without degradation
- The ability to process signals for security and other purposes
- The ability to remove unnecessary redundancy to increase transmission and storage efficiency
- The ability to add appropriate coding and redundancy so that signals can survive harsh transmission environments

3.2.1 High-Definition Television (HDTV)

The big push for consumer digital television came as a consequence of the search for a practical high-definition television (HDTV) system. HDTV was defined as having twice the horizontal resolution, twice the vertical resolution, a wide picture with a ratio of 16 units of width for every 9 units of height, no visible artifacts at reasonable viewing distances, and compact disc–quality sound. Before any processing, the analog signal coming from an HDTV camera could consist of 30 MHz of red information, 30 MHz of green information, and 30 MHz of blue information. Almost 100 MHz of analog information is involved. To convert it to digital form, it is first sampled at twice the highest frequency (according to the Nyquist theorem; see Section 3.3.2). Then each sample is represented by a byte of data. More than a gigabit per second of data transmission is required. It can be appreciated why all the early HDTV proposals were analog! Methods of reducing this bandwidth appetite had to be found. The work on HDTV began in Japan in 1970 using analog transmission technology coupled with significant digital processing at both the point of origination and the receiver. The Japanese proposal was called multiple sub-Nyquist sampling encoding (MUSE). It applied many of the techniques used to minimize NTSC bandwidth. The goal was to match the HDTV system to the human visual response, since there is no need to transmit what the eye does not see.

When the FCC decided to pursue HDTV in the United States, it laid down a number of challenges.[3] The first challenge was for a "compatible" HDTV system that would not make current NTSC receivers obsolete. The second major challenge was to limit the original NTSC and the new HDTV signals to no more than two 6-MHz slots in the television spectrum. A further restriction required the new signals to avoid undue harm to existing NTSC transmissions. After an extended search for a compatible method of creating HDTV, it became clear that all methods proposed used the original NTSC signal plus in-band and out-of-band "helper signals." All these resources were required to create the compatible signal, and two 6-MHz bands were consumed for each HDTV signal. This approach meant that NTSC would always have to be supported. It also meant that 12 MHz was committed for each HDTV signal. If there were ever to be a transition away from broadcast NTSC, this approach would have to be abandoned.

Zenith Electronics Corporation first broke ranks with the analog proponents by proposing a hybrid system that continued the transmission of the high frequencies of the image in analog form but converted the lower frequencies to a digitized form. This hybrid approach seemed to use the best of both worlds. It recognized that most of the energy in an NTSC signal is in its low frequencies, which include the synchronization pulses. In NTSC, the sync pulses have the highest energy because they are separated from the video signal by amplitude discrimination. By digitizing the low frequencies, the majority of their power consumption was eliminated, yet the burden on the digital circuits was relaxed because only relatively low frequencies were processed. The high frequencies

remained analog and contributed little to the power requirements. The lower-data-rate digital signals would also be less susceptible to multipath transmission effects. The remaining problem was that this approach was no longer compatible with existing NTSC receivers. This problem was solved by allowing compatibility to include simulcasting. That is, both the hybrid signal and the NTSC signal would carry the same programming, but at two different resolutions and at two different frequencies. This would preserve the utility of older receivers. Moreover, since no successful system that put both NTSC and HDTV into the same 6 MHz had been proposed, two 6-MHz channels would still be required, as in all the other proposed systems.

This approach had a number of major advantages. The low-power HDTV signal was tailored to cause minimum co-channel interference in adjacent locations using the same frequencies. Also, since the NTSC signal was not needed as a component of the HDTV signal, it could eventually be abandoned. The NTSC channel could then be reallocated to other purposes. Even before that happened, the requirement for simulcasting could be relaxed based on policy rather than technological constraints. Via this step-by-step process, compatibility was abandoned for the first time in broadcast television. (The noncompatible CBS color system, though temporarily the official system, did not achieve commercial success before it was replaced with the RCA/NTSC compatible color system.)

However, cable and satellite didn't have the same requirement for terrestrial compatibility. In fact, the lack of any viable analog HDTV solution for satellite led General Instrument's DigiCipher division to develop their purely digital approach. (At the same time the "clean sheet of paper" approach was also being advocated in some cable circles.) Shortly thereafter, General Instrument proposed an all-digital solution for terrestrial HDTV broadcast. Quickly, all serious proponents, with the exception of the Japanese MUSE system, converted to all-digital designs. The committee, charged with selecting a winner, found that it could not come to a conclusion. The technical issues were too complex, and the political issues were overwhelming. At the time a decision was to be made, all the proposed systems made unacceptable pictures. The result was an elaborate ruse to score all the systems as acceptable under the condition that a "grand alliance" be formed that allowed the proponents themselves to hammer out a single system. This allowed the political battles to go on behind closed doors under the guise of selecting the "best of the best" for a single proposal to the FCC.

Since that time, the television and computer industries have jointly created standardized toolkits for audio and video compression. The MPEG-2 video compression standard[4] provides a "toolbox" of techniques that can be selected according to the nature of an application. The MPEG-2 systems standard defines how multiple audio and video channels can be multiplexed together into a single digital bitstream. Finally, the broadcast and cable industries have developed modulation and transmission standards to reliably deliver the digital bitstream over an analog channel.

Using these standards, a single HDTV program can now be transmitted within the analog broadcast TV channel assignment of 6 MHz rather than the tens of megahertz once thought necessary. In the case of cable's well-behaved spectrum, double the data transmission rate is possible; two to three HDTV signals can be carried in 6 MHz.

The same technology that makes HDTV possible in 12–18 Mbps also enables very efficient compression of standard-definition television. Excellent results are now being obtained at data rates of 2–3 Mb/s using MPEG-2 video compression. Since the cable transmission rate is approximately 38 Mb/s (in 6 MHz), 12–18 programs can be carried in the same spectrum that previously carried one.

It is expected that the transmission of standard analog NTSC will occur for many years before a transition to digital television is completed. The British experience with vertically polarized, 405-line monochrome television is instructive. It took some 50 years to abandon that system, even with a very small number of receivers. A minority of opinion holds that the U.S. transition may never be completed because of the vast installed base of analog receivers (currently more than 250 million color receivers and more than 150 million VCRs) and the political problems associated with making consumers' "investment" in products obsolete. Low-income viewers and the economically challenged would be especially hard hit. It is difficult simultaneously to be a supporter of "free television" and to require the poor to purchase new, expensive television sets (not to mention that advertiser-supported television requires a large number of subscribers to sustain itself). Of course, the new digital receivers will rapidly come down in price, but the analog receivers will also continue to decline in price. The shutdown of analog television will not be easy.

3.2.2 Digital Signal Processing

The human sensory systems for images and sounds are analog. Images and sounds start as analog phenomena. To be enjoyed by humans, the phenomena must be converted to signals, and the signals must eventually be displayed as analog stimuli for eyes and ears to enjoy. Unfortunately, as analog signals are transmitted over long distances, they encounter noise, distortion, and interfering signals that degrade the quality of the images and sounds, so they become unpleasant and, eventually, unusable. If the analog signals are converted to digital signals, an arbitrarily small degree of noise is introduced in the conversion process, but all subsequent degradation of the signal can be avoided using practical and well-understood techniques.

The advantages of baseband (unmodulated) digital signals include the ability to completely regenerate the signal, thereby preventing the accumulation of noise and distortion, and the ability to apply computational techniques for the purposes of error detection and correction and redundancy reduction. Redundancy reduction is important because it saves data storage space and transmission time and bandwidth.

Advantages of binary numbers include that they have two states, which can be represented by simple and inexpensive circuits, and that the impact of electrical noise and distortion can be minimized or even eliminated by proper design. A circuit element, such as a transistor, that processes an analog signal must faithfully reproduce all values of the signal while adding a minimum of distortion and noise. If many such circuit elements process an analog signal, their individual contributions of noise and distortion accumulate, causing substantial signal degradation. On the other hand, a binary circuit can have two well-defined states, on and off, which are easily distinguishable. The on state may represent the binary number 1, whereas the off state can represent the binary number 0. (The opposite choice is equally valid.) The important point is that if the circuit element is *mostly off* but not completely off, it will not be confused with the on state. Similarly, if the circuit element is *mostly on* but not completely on, it will not be confused with the off state. Thus imperfect performance of the circuit can still faithfully represent the binary values. Only when the on state approaches half of the assigned value or when the off state is almost halfway to the on condition can confusion result. If this degree of deficient performance is avoided, the two states can be discriminated and the signal perfectly resolved (see Figure 3.1). If, as the signal is transmitted, it suffers some noise and distortion degradation, it can still be perfectly recovered as long as the two states — the on state representing a binary 1 and the off state representing a binary 0 — can be reliably discriminated. Eventually, sufficient noise and distortion will accumulate so that the two states become confused. If the system is designed so that the signal is regenerated (and any errors are corrected) prior to this destructive level of degradation, a fresh binary signal can be substituted for the degraded signal, and all the damage caused by noise and distortion can be completely undone. *Repeaters* can perform this operation an arbitrary number of times, allowing error-free communications over arbitrarily long distances. This is something that cannot be accomplished with analog signals.

A further advantage of digital signals is that the art of circuit design has progressed to where the transistors that process digital signals are very small and very inexpensive. Gordon Moore, one of the founders of Intel Corporation, observed that approximately every 12–18 months, the number of digital transistors that can be afforded for a given price doubles. Alternatively, the cost of a given number of digital transistors approximately halves in that period. This process has been continuing for decades and appears likely to continue for some time to come. As an example of this phenomenon, the first personal computers introduced in the early 1980s used an Intel-brand integrated circuit (IC) that had 30,000 digital transistors. The Pentium computer ICs of the mid-1990s have more than 5 million digital transistors. Tens of millions of digital transistors are now being placed in consumer products at affordable prices. The same experience has not been enjoyed by analog circuits, because they must faithfully process the infinite range of values of analog signals. That severe constraint has prevented analog circuits from progressing as fast or as far in complexity and cost reduction.

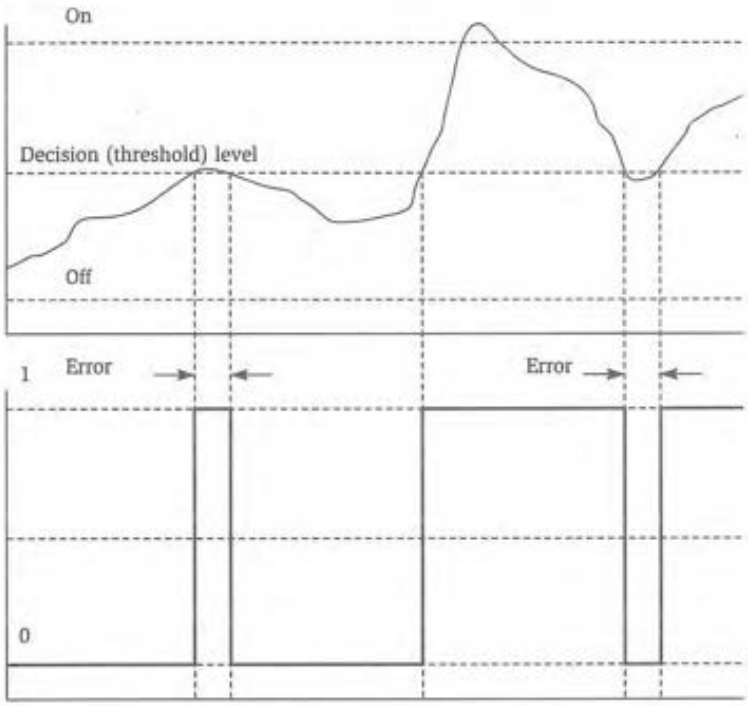


Figure 3.1 Impact of excessive noise.

An additional advantage of digital signals is that they can be mathematically manipulated in very complex ways that lead to methods of determining whether transmission errors have occurred and even of correcting some of those errors. Note that there are only two possible types of errors: a binary 1 symbol may be damaged and converted into a binary 0 symbol, and a binary 0 symbol may be damaged and converted into a binary 1 symbol. There are no other alternatives in a binary system. As an example, one common method of identifying an error condition is called *parity detection*. Binary symbols are grouped into clusters of seven, and an eighth symbol is appended depending on whether the previous seven have an even or an odd number of 1 symbols. If the appended symbol is such that an even number of 1 symbols is always present in each valid group of eight symbols, the method is called *even parity*. A single transmission error will convert a 1 to a 0 or a 0 to a 1, and the result will be an odd number of 1 symbols. The system can detect this error. If two errors occur, a much less probable event, the system will be fooled and think that no error has occurred, because there will still be an even number of 1s. However, if three errors occur, the damage will again be detected. The ability to detect certain error conditions has come at the

price of an appended symbol that takes up transmission time and that requires additional circuits to process the signal at both the transmission and receiving ends. Forward error correction (FEC) algorithms can detect multiple errors and even compute the original correct signal. These more complex methods increase the amount of additional, nondata symbols and are said to have increased *overhead*. In addition, additional processing complexity is required at both the sending and receiving ends of the transmission path.

3.3 Digital Video Compression

As we have seen in Section 3.2, uncompressed digital video requires a very high data rate; in fact, these rates are so high that it would be uneconomical to transmit or store uncompressed digital video. Therefore data compression is required so that the video is still of acceptable quality while being described by many fewer bits. This type of compression is known as *lossy* compression, because some information is discarded during the encoding process. "Acceptable quality" is defined by customer expectations and business conditions and is, by nature, always something of a compromise. For both operators and programmers, more is better; the more aggressive the compression, the more channels for a given bandwidth. Many television viewers seem to be very forgiving of a softer picture (that is, a loss of resolution) as long as the picture remains noise free. Nevertheless, more sophisticated viewers have certainly shown their preference for better resolution, as is demonstrated by the phenomenal success of DVD. This trend may continue as large-screen, high-resolution displays become more affordable, and it bodes well for HDTV adoption.

The first large-scale application of digital video compression was, and still is, broadcast television services (in the order of satellite, cable, and terrestrial broadcast). These applications all require a standard coding mechanism that is understood by all decoders in the marketplace. The familiar MPEG standard is the best example of this. It also allowed for volume production of MPEG-2 decoder chips so that they became inexpensive enough to put in the set-top terminal (STT) and, eventually, the television set.

A second application is the digital versatile (originally video) disk (DVD). Just as in broadcast television services, the decoder in every DVD player had to perform identically; once again, the MPEG-2 video format was adopted.

Finally, video on demand (VOD), which is now being aggressively deployed by many cable operators, also uses MPEG-2 standards. In this case, MPEG-2 is a natural choice, because VOD leverages the existing investment in the STT.

A very different application, called *streaming media*, has emerged with the Internet. In streaming media, the transfer of video is typically a 1-to-1 transaction, making it analogous to VOD. However, the Internet environment allows the coding format to be specified for each transaction, even allowing a software decoder to be downloaded just before it is needed. Moreover, the

ubiquitous personal computer now provides enough raw central processing unit (CPU) performance and memory to allow the decoding to be done without special-purpose silicon. These properties make the Internet a perfect incubator for new compression algorithms.

3.3.1 Principles of Video Compression

There are many excellent texts that describe the principles of digital video compression.[5-7] The reader is referred to these for a more detailed treatment.

Digital video compression can achieve a dramatic reduction in bit rate; a factor of 50:1 is quite achievable for entertainment-quality video. There are two main reasons for this.

1. A typical video stream contains a tremendous amount of spatial and temporal redundancy.
2. The human eye–brain system is relatively insensitive to certain impairments in the displayed result.

In the first case, recall that NTSC video is refreshed 60 times per second (and PAL video 50 times per second) to avoid the perception of flicker. There are many segments when very little motion has to be conveyed, and one field is much like the one before it (temporal redundancy). Fundamentally, if the changes could be sent, rather than completely refreshing the field, then a lot of data transmission could be saved.

Second, the human eye–brain system is quite forgiving when it comes to the fidelity of video rendition. For example, the eye's rod cells are very sensitive to the total intensity of light (luminance), whereas the eye's cone cells are less sensitive to color differences, particularly for certain shades. The eye is also less sensitive to color (or chrominance) detail. NTSC takes advantage of this to dramatically limit the bandwidth of the chrominance (I and Q) signals, as explained in Chapter 2. Finally, the eye–brain system is even less sensitive to detail if there is significant motion in that part of the scene.

These discoveries about the human eye–brain system have been researched thoroughly and codified into a perceptual model that can be used to reduce fidelity in those areas where it is least noticeable while maintaining fidelity where the eye is most discriminating. Nevertheless, much advanced compression work transcends perceptual modeling, becoming quite empirical. This explains the lack of good, objective video-quality measurement techniques based on perceptual modeling.

3.3.2 Coding of Still Images

We will start by discussing the compression of a still image and then extend the model to deal with moving images. Video can be represented as a series of still

images. In fact, we must periodically code an entire frame as a still image (an intracoded frame, or I frame) because this provides the property of random access. An I frame contains all of the information needed to decode the frame within the frame itself; it is not dependent on adjacent frames, and it serves as a reference for these other frame types. Random access is very important for broadcast applications; when the viewer is switching from one channel to the next, it must be possible to reinitialize the decoding process within a short period of time (less than a second).

In both still- and moving-image compression, a number of key steps are commonly applied:

1. Sampling and preprocessing
2. Block formation
3. Transform coding
4. Quantization
5. Run-length encoding
6. Entropy coding

(This section describes how the fundamental principles of compression are implemented by MPEG-2 video compression of a standard-definition TV picture, because it is so widely used. Subsequent sections will expand the scope to include discussion of HDTV, MPEG-4 video, and alternative approaches to video encoding.)

Sampling and Preprocessing

To convert an analog signal to a digital signal, it is first sampled in time. This can be accomplished by multiplying the analog waveform by a train of evenly spaced narrow pulses that alternate between zero volts and a constant, convenient voltage (for example, 1 volt) (see Figure 3.2). The information science theorist Harry Nyquist of Bell Telephone Laboratories proved that if a signal is sampled at least twice as frequently as the maximum signal frequency it contains, it can be perfectly recovered with no loss of information.[8,9] This principle is called the Nyquist theorem. Sampled signals are still analog signals, because they can take on any value; they are just time quantized.

Nyquist proved that if the sampled signal is low-pass filtered, the original signal is recovered, with just a delay caused by the propagation time through the circuits (see Figure 3.3). If each time sample's strength is then measured (for example, with a digital voltmeter) and the resulting measurement represented by a number of limited precision, the sampled analog signal will be converted into a sequence of data (see Figure 3.4). Limited-precision numbers have a fixed number of decimal places. The uncertainty in precision of the number is determined by the value of its last decimal place. The information to be transmitted is

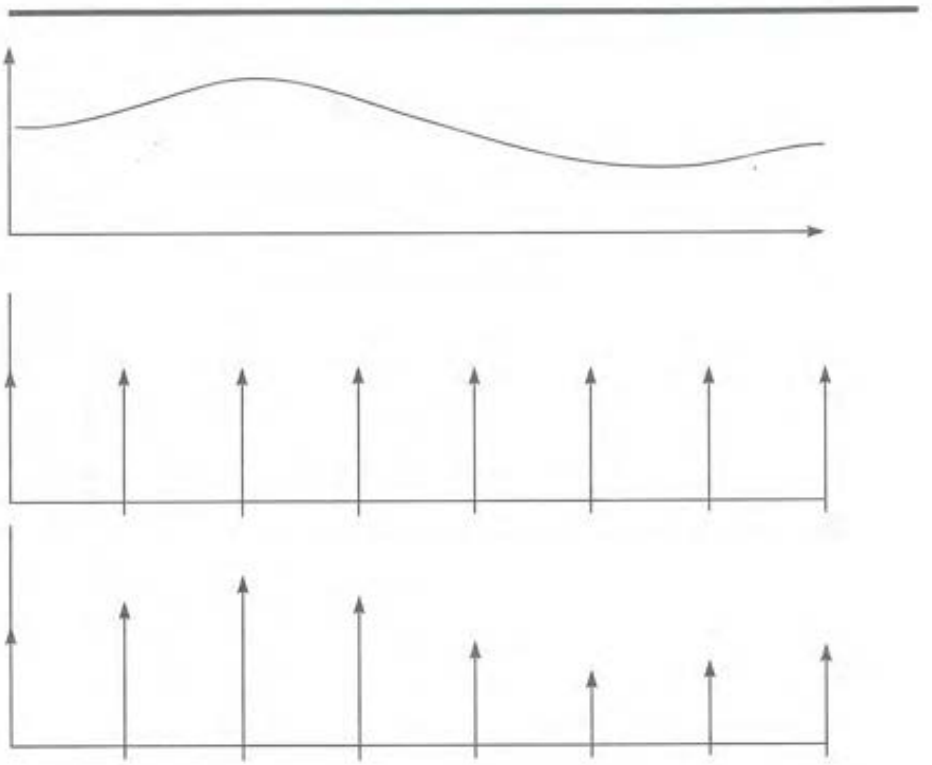


Figure 3.2 Time sampling an analog waveform.

no longer the original analog signal or its time-sampled version (which can take on any value), but rather another signal that conveys the limited-precision numbers describing the strength of the original signal samples. The representation of the signal by a limited-precision number introduces an error that can be considered as a type of noise called *quantization noise*. The amount of quantization noise can be made arbitrarily small by using arbitrarily higher-precision numbers, but it can never be reduced to zero.

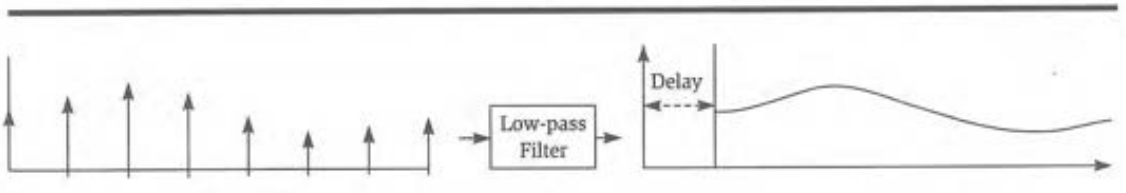


Figure 3.3 Recovery of a time-sampled waveform.

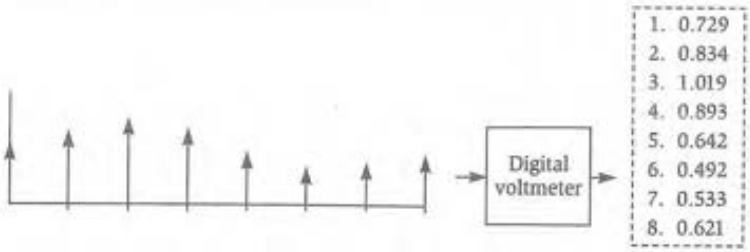


Figure 3.4 Quantization of a time-sampled waveform.

Video can be sampled in many different ways—ironically, sampling to achieve studio-quality transfers also tends to generate the highest data rate and is therefore not the best approach for low-bit-rate encoding. However, the "cleaner" the video, the better, because any noise in the original signal is random, additional information that is not easy to compress. For this reason, it is common practice to low-pass filter the video signals before sampling, in an attempt to reduce any high-frequency noise components. Achieving good noise reduction without unduly reducing resolution has been a subject of considerable research and development among encoder manufacturers.

As described in the previous chapter, color television is based on the discovery that the human eye responds to three colors and mixes them to produce essentially all other colors. Color television requires three signals. The three signals start at the camera and end at the picture tube as red, green, and blue. In between, they can be transformed into other formats that are more convenient for processing. In digital television, the color information is sent as two color difference signals, or *components*. The first component, C_b, is the difference between the luminance value and blue in that part of the picture. The second component, C_r, is the difference between the luminance value and red in that part of the picture. Since the human eye is less sensitive to detail in color information than in luminance (brightness), the color information can be sent with less detail.

As mentioned already, to be properly reproduced, an analog signal must be sampled at a rate at least twice that of its highest frequency. That theoretically minimum rate requires perfect (nonrealizable) filters at the receiving end. To use practical filters, an even higher sampling rate is required. The most common sampling rates chosen for use in digital television with resolution comparable to NTSC are 13.5 MHz for the luminance signal and half that rate (6.75 MHz) for each of the two components, for a total of 27 megasamples per second.

Interestingly, the use of compression allows the program bandwidth and the transmitted bandwidth to be very different. A modern television camera for NTSC actually produces video frequencies beyond what can be transmitted in the 6-MHz broadcast channel. Those frequencies are just thrown away in analog

practice. In compressed television, some of them may be used. The 13.5-MHz sampling rate supports a luminance bandwidth of 5.75 MHz and supports 2.75 MHz in each of the color difference channels. The NTSC technique of choosing color axes that favor the human visual system and limiting one to 0.5 MHz while allowing the other to go to 1.5 MHz is not necessary. Compression allows both color difference signals to have the same 2.75-MHz bandwidth.

It was determined by experiment that a minimum of eight bits per sample are required to minimize visible artifacts due to the quantization of the data. In some applications, 10 bits per sample are used. In digital television systems using 27 megasamples per second and eight bits per sample, uncompressed transmission requires 215 Mb/s for the video data. Of course, audio must be digitized as well. Using our bandwidth rule of thumb, something over 100 MHz of bandwidth would be required. This is not a problem internal to systems that process a single channel or for transmission of a few channels over fiber-optic links. However, this kind of bandwidth consumption is not feasible for broadcast or for multichannel environments, where the number of channels is important to subscriber appeal. There is a clear need for methods of reducing the bit rate to acceptable levels.

Because NTSC has fewer lines (525) and a higher frame rate (29.97), while PAL has more lines (625) and a lower frame rate (25), both forms of analog video can be conveniently sampled at the same rate to create the basis for a worldwide digital video rate hierarchy.[10] Sampling of an analog video source according to ITU-T Recommendation BT.601-5[11] reduces the chrominance data rate by sampling at 4:2:2.* The meaning of these numbers is:

4. The luminance component (Y) is sampled at 13.5 MHz.
2. The first chrominance component (C_r) is sampled at 6.75 MHz.
2. The second chrominance component (C_b) is sampled at 6.75 MHz.

Here, Y, C_b,C_r are the digital equivalents of the luminance, I, and Q components of an analog picture. Just as in the analog case, the color difference signals are manipulated to provide the best fidelity where it is most critical for the human eye.

In both NTSC and PAL, the active picture region is 720 pixels wide, but the first eight and the last eight are usually ignored, so only the middle 704 pixels are encoded. In the vertical direction, after VBI lines are discarded, there are 483 active lines in an NTSC picture, though only 480 lines are encoded. For PAL, depending on the version, there are between 575 and 587 active lines; only 576 are encoded.

* Originally the number 4 originated as being 4 times the color subcarrier frequency — for NTSC this is 4 × 3.58 MHz (or 14.3 MHz); however, a slightly lower rate of 13.5 MHz was chosen by ITU-R BT.601-5 for compatibility with PAL.

Block Formation

The next step in the process is to represent the video as a two-dimensional array of pixels so that it can be processed in the spatial domain. To do this, the digital video stream is deinterlaced and clocked into a frame buffer. (For images with high-motion content, this can cause a double image because there is movement between the two frames. We'll discuss ways to deal with this problem later; see Section 3.3.3, Coding of Moving Images.)

The frame buffer is divided into blocks typically 8 × 8 pixels because this is a convenient size for transform coding (see next subsection). Before this is done, however, the chrominance component is subsampled again, this time in the vertical direction, by averaging the samples on adjacent rows in the frame buffer. Sampling at 6.75 MHz and then averaging between rows reduces the chrominance resolution by a factor of 4 (compared to luminance). This resolution is termed 4:2:0. Therefore, for every four Y (luminance) blocks, there is one C_b block and one C_r block, as shown in Figure 3.5. (This grouping of blocks is called a macroblock in MPEG parlance.)

The vital statistics for standard-definition digital video are summarized in Table 3.1.

Transform Coding

As discussed in the previous chapter, waveforms that are cyclically repetitive can be decomposed into a series of sine and cosine waves of varying amplitudes. Equations (3.1), (3.2), and (3.3) display the familiar Fourier series equations discussed in the last chapter.[12]

$$f(t) = a_0 + \sum_{n=1}^{\infty}[a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)] \qquad (t_0 < t < t_0 + T) \qquad (3.1)$$

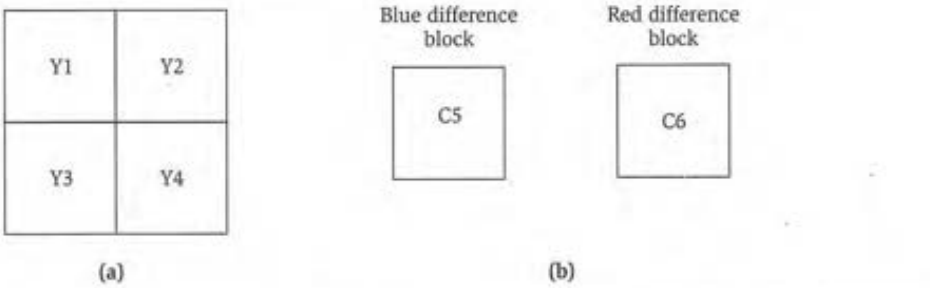


Figure 3.5 Order of transmission. (a) Luminance blocks. (b) Color difference blocks.

Table 3.1 Vital Statistics for Standard-Definition Digital Video

ITU-T BT.601-5/MPEG-2 main level	*NTSC*	*PAL*
Frame rate	29.97 Hz	25 Hz
Sampling rate:		
Luminance	13.5 MHz	13.5 MHz
Chrominance	6.75 MHz	6.75 MHz
Samples per line	858	864
Active picture samples	720	720
Encoded pixels per line	704	720
Vertical lines	525	625
Active lines	483	575–587
Encoded lines	480	576
Frame resolution (pixels)	704 × 480	704 × 576
Total macroblocks per frame	1,320	1,584
Total blocks per frame	7,920	9,504

Equation (3.1) is the trigonometric Fourier series representation of time waveform $f(t)$ over time interval $(t_0, t_0 + T)$. The various constants a_n and b_n are given by

$$a_n = \frac{\int_{t_0}^{(t_0+T)} f(t)\cos(n\omega_0 t)dt}{\int_{t_0}^{(t_0+T)} \cos^2(n\omega_0 t)dt} \tag{3.2}$$

and

$$b_n = \frac{\int_{t_0}^{(t_0+T)} f(t)\sin(n\omega_0 t)dt}{\int_{t_0}^{(t_0+T)} \sin^2(n\omega_0 t)dt} \tag{3.3}$$

If a waveform is not periodic, such as shown in Figure 3.6(a), it is possible to define a periodic version, as shown in Figure 3.6(b).

The Fourier series in Equations (3.1), (3.2), and (3.3) will be valid over the interval from time t_0 to time $t_0 + T$ but incorrect elsewhere. This is not a problem if we simply reapply the technique of forming periodic extensions for each piece of the waveform. We can then process the signal one piece at a time and load the results in memory.

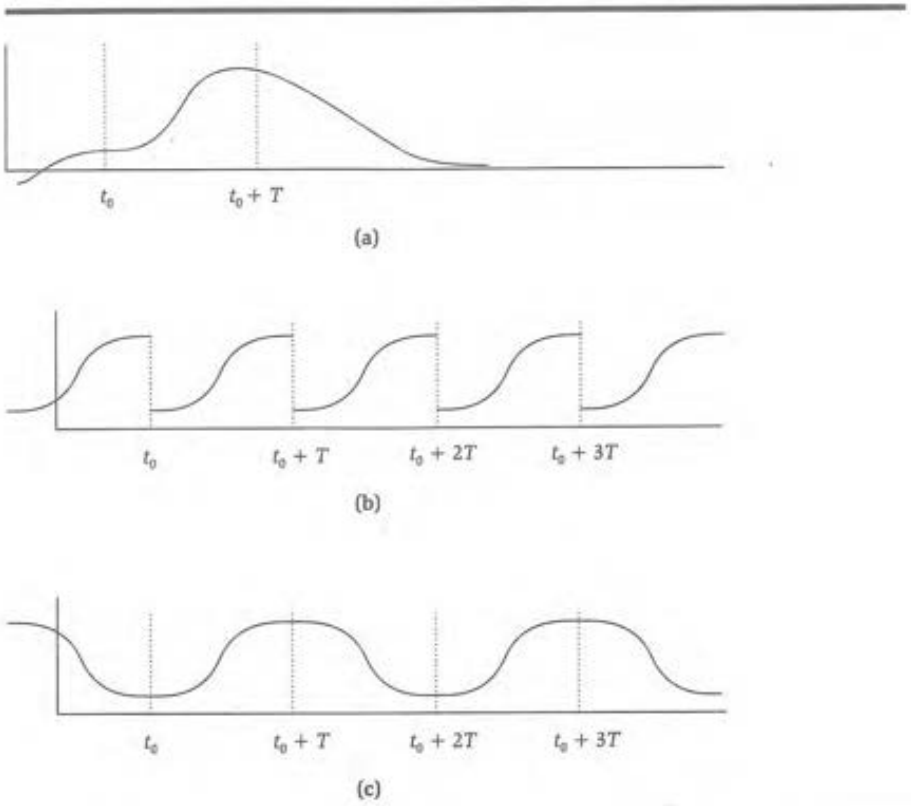


Figure 3.6 Periodic construction. (a) Nonperiodic waveform. (b) Periodic extension. (c) Symmetric periodic extension.

If a waveform has symmetry about the zero time point, then only cosines are needed to represent it, and all the b_n terms become zero. These cosines occur at multiples — or harmonics — of the fundamental repeat frequency of the waveform.

Figure 3.6(c) shows that the periodic extension technique can be applied in a manner that makes the signal symmetric so that its b_n terms will become zero. Again, the Fourier series is valid over only the interval from time t_0 to time $t_0 + T$ but incorrect elsewhere.

The amplitude values of these cosine waves constitute the cosine Fourier series representation of the waveform. The Fourier representation simplifies to just Equations (3.4) and (3.5).

$$f(t) = a_0 + \sum_{n=1}^{\infty}(a_n \cos n\omega_0 t) \qquad (t_0 < t < t_0 + T) \tag{3.4}$$

$$a_n = \frac{\int_{t_0}^{(t_0+T)} f(t)\cos(n\omega_0 t)dt}{\int_{t_0}^{(t_0+T)} \cos^2(n\omega_0 t)dt} \tag{3.5}$$

We are going to sample the video waveform according to the Nyquist theorem (as discussed in Section 3.3.2). Samples will be taken at more than twice the video waveform's highest frequency. Consequently, the video waveform will no longer be continuous. It will exist only at a finite number of discrete points in time. Therefore the integrals of Equation (3.5) become simple summations because there is only a finite number of sample points to add up rather than a continuous set of values to integrate. Since we are dealing with a bandwidth-limited signal, the summation of Equation (3.4) does not go to infinity, just to the highest frequency in the band-limited signal. In this way, the cosine Fourier series becomes a discrete cosine transform (DCT).

A picture is a two-dimensional object. The same principles that applied in the horizontal direction of the graph of Figure 3.6(a) can be applied in the vertical direction of a picture as well. Instead of the time variable in Figure 3.6, a distance variable is used. This is equally valid. (Of course, distance along a raster-scanned image has a relationship to time!) These principles also apply to both directions simultaneously. To apply the DCT, we need a periodic extension of the image. This is best done by breaking the image up into smaller blocks. The size of block chosen in most digital television applications is 8 picture elements (pixels) wide and 8 pixels high. This block can be imagined to repeat in a checkerboard pattern so that it is a symmetrical replication from checkerboard square to checkerboard square. This allows us to utilize only cosines and avoid the sines entirely. However, just as with the time waveform, this DCT will be valid only within the chosen block. Each block of the image will have to be handled separately. Again, bandwidth limiting of the video means that the image details are limited in the number of variations they can make within the block. This means that the DCT summations have a fixed maximum number and don't have to go to infinity.

The video samples in two directions are transformed into a form of spectrum by the two-dimensional DCT of Equations (3.6), (3.7), and (3.8)[13]:

$$F(u,v) = \frac{1}{4}C(u)C(v)\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y)\cos\left[\frac{(2x+1)u\pi}{16}\right]\cos\left[\frac{(2y+1)v\pi}{16}\right] \tag{3.6}$$

where

x and y = pixel indices within an 8-by-8 video block

That is, x and y are the row and column numbers of the block that contains the measurements of the picture's brightness at each location in the block. Similarly,

u and v are DCT coefficient indices within an 8-by-8 spectrum block. That is, u and v are the row and column numbers of the block that contains the measurements of the picture's spectrum at each location in the block. Here,

$$C(u) = \frac{1}{\sqrt{2}} \quad \text{for } u = 0 \quad \text{and} \quad C(u) = 1 \quad \text{for all other values of } u \tag{3.7}$$

$$C(v) = \frac{1}{\sqrt{2}} \quad \text{for } v = 0 \quad \text{and} \quad C(v) = 1 \quad \text{for all other values of } v \tag{3.8}$$

To summarize, when we take a block 8 pixels wide and 8 pixels high from a bandwidth-limited video signal, there are 64 pixels. Each pixel is sampled in time. Its amplitude is measured, and the measurement is rounded to the nearest quantizing level. This quantizes the amplitude to a resolution of 8 bits (256 levels). We then have an array of 64 numbers in an 8-by-8 square (see Figure 3.7(a)).

If we apply Equations (3.6), (3.7), and (3.8) to this array, we will get another array of 64 numbers in 8 rows and 8 columns (see Figure 3.7(b)). This array represents the frequency components of the original amplitude array. The cells in this array contain numbers called the *frequency coefficients*. The number in the upper left-hand corner is related to the direct current (dc) component, that is, the average amplitude across the video square. The cells across the top row are the values of the horizontal frequencies in the block. The cells down the left side of the block are the vertical frequencies in the block. The other blocks represent frequencies in other directions (see Figure 3.7(c)).

The Fourier series works in both directions. If we have an amplitude waveform, we can convert it into a frequency series. If we have a frequency series, we can convert it back to the waveform. Similarly, there is an inverse discrete cosine transform (IDCT), which will convert the frequency coefficients array back into the amplitude array. Here are the equations for the IDCT[14]:

$$f(x,y) = \frac{1}{4}\sum_{u=0}^{7}\sum_{v=0}^{7} C(u)C(v)F(u,v)\cos\left[\frac{(2x+1)u\pi}{16}\right]\cos\left[\frac{(2y+1)v\pi}{16}\right] \tag{3.9}$$

$$C(x) = \frac{1}{\sqrt{2}} \quad \text{for } x = 0 \quad \text{and} \quad C(x) = 1 \quad \text{for all other values of } x \tag{3.10}$$

$$C(y) = \frac{1}{\sqrt{2}} \quad \text{for } y = 0 \quad \text{and} \quad C(y) = 1 \quad \text{for all other values of } y \tag{3.11}$$

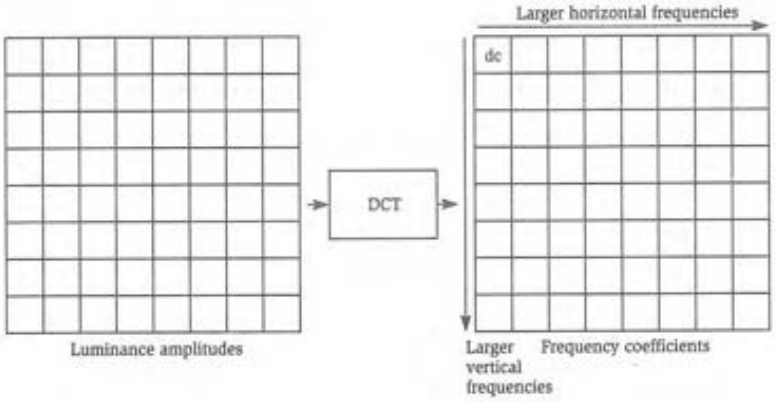
Applying these equations to the array of Figure 3.7(b) yields the array of Figure 3.8, which is recognized as the original amplitude array. This is a lossless conversion (to the limits of computational accuracy).

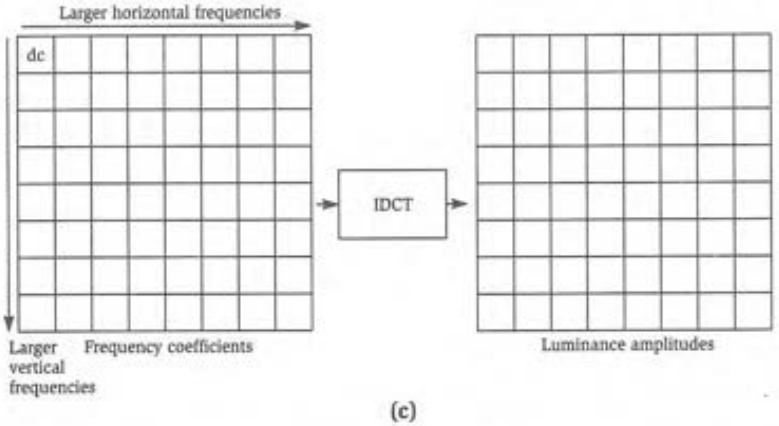
139	144	149	153	155	155	155	155
144	151	153	156	159	156	156	156
150	155	160	163	158	156	156	156
159	161	162	160	160	159	159	159
159	160	161	162	162	155	155	155
161	161	161	161	160	157	157	157
162	162	161	163	162	157	157	157
162	162	161	161	163	158	158	158

(a)

235.6	−1	−12.1	−5.2	2.1	−1.7	−2.7	1.3
−22.6	−17.5	−6.2	−3.2	−2.9	−0.1	0.4	−1.2
−10.9	−9.3	−1.6	1.5	0.2	−0.9	−0.6	−0.1
−7.1	−1.9	0.2	1.5	0.9	−0.1	0	0.3
−0.6	−0.8	1.5	1.6	−0.1	−0.7	0.6	1.3
1.8	−0.2	1.6	−0.3	−0.8	1.5	1	−1
−1.3	−0.4	−0.3	−1.5	−0.5	1.7	1.1	−0.8
−2.6	1.6	−3.8	−1.8	1.9	1.2	−0.6	−0.4

(b)





(c)

Figure 3.7 DCT 8-by-8 blocks. (a) Luminance values. (b) Frequency coefficients after DCT. (c) DCT and IDCT transformations.

139	144	149	153	155	155	155	155
144	151	153	156	159	156	156	156
150	155	160	163	158	156	156	156
159	161	162	160	160	159	159	159
159	160	161	162	162	155	155	155
161	161	161	161	160	157	157	157
162	162	161	163	162	157	157	157
162	162	161	161	163	158	158	158

Figure 3.8 Luminance values after IDCT.

The order of transmission of this information is shown in Figure 3.5. First, four blocks of luminance information are sent. Then one block of each of the two color difference signals is sent. The color difference signals come from sampling every other location in the luminance array of four blocks. All six of these blocks are separately converted by the DCT circuits and separately transmitted to the receiver.

Quantization

Let's review what has been accomplished so far: The video signal has been time sampled and quantized in a manner that gives the values of the luminance and color difference components at specific points in the picture. Those luminance points have been organized into blocks 8 pixels wide by 8 pixels high. The same has been done for the color difference signals, except every other pixel has been skipped to form macroblocks that span 16-by-16 luminance pixels. Then the values of these pixels have been converted to frequency values using the DCT in a manner that allows the original values to be computed again using the IDCT. This digitization and conversion by the DCT has not, by itself, done anything to reduce the redundancy of the image. More processing is needed to eliminate redundancy.[15]

The first step in the process is to note that there is a much wider range of values in the frequency domain cells than in the original video domain cells. Many of the frequency numbers are very small. In order to efficiently transmit the information in the frequency cells, we must reduce the amount of information necessary to convey it. It is already sampled and digitized, so the next step is to quantize the values in the frequency cells.

The transform coefficients represent different frequencies in brightness and color. Studies have shown that the human visual system has differing sensitivities to this information, depending on the amount of motion, the amount of detail, and the color of the image. There is a lot more information in this data than the human visual system requires for a satisfactory viewing experience. Consequently, the values in the cells can be approximated by a smaller set of values. The actual cell value is compared with the smaller set, and the nearest set value is chosen to represent the actual value. This process is another quantization mechanism. The cells are quantized differently according to the results of the research on human perception. A *quantizer matrix* is assigned to the transform coefficient matrix that prescribes the coarseness of quantization for each cell. Predetermined quantizer matrices are built into the system. Additionally, specific quantizer matrices can be downloaded as part of the signal for special cases.

Coarser quantization requires fewer bits, and the video matrix consisting of 64 cells with 8 bits in each row is transformed into 64 cells with fewer total bits. A most important result is that many of the high-frequency blocks contain coefficients that are so small that they become zero when quantized to a coarse scale. Thus the coefficient matrix is loaded with zeros. In addition, a quantizer coefficient is applied to the entire quantizer matrix to adjust its coarseness relative to other blocks in the picture. In this way, some areas of the picture are assigned more bits than other areas in relation to their importance to the image.

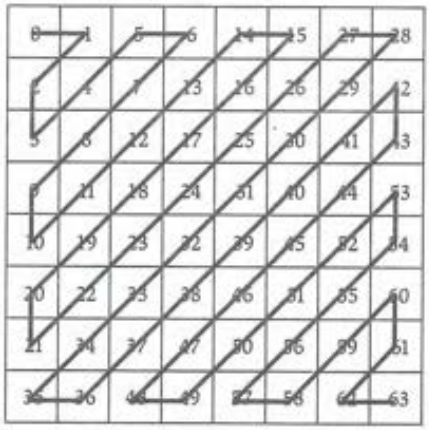
Some of the image's blocks will have little detail and many zeros in their coefficient matrices and thus will use few bits. Other blocks will have a great deal of detail and relatively few zeros in their coefficient matrices and will require many more bits. In this manner, bits are allocated to the parts of the picture that need them.

Run Length Encoding

The order of transmission of the values in the coefficient matrix is very important. A process called *zigzag scanning* is employed. Figure 3.9(a) illustrates this. In Figure 3.9(b), the frequency coefficients from Figure 3.7(b) are plotted as a graph in the order scanned in Figure 3.9(a). (In Figure 3.9(c), the large direct current (dc) term is not plotted so that the relative sizes of the remaining terms can be better seen.) The frequency coefficients quickly become small in value as the scanning progresses through the block. Most can be rounded to zeros, yielding a "run of zeros." Rather than transmit all the zeros, a pair of numbers is transmitted. The first number of the pair tells the number of zeros. The second number tells the value of the cell that terminates the run of zeros.[16] When a condition is reached where all the remaining cells have zeros, an end-of-block code is sent. Already, the number of bits to be transmitted has been significantly reduced.

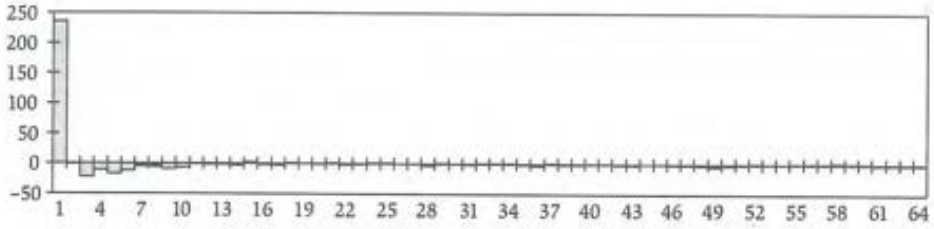
0	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Luminance amplitudes

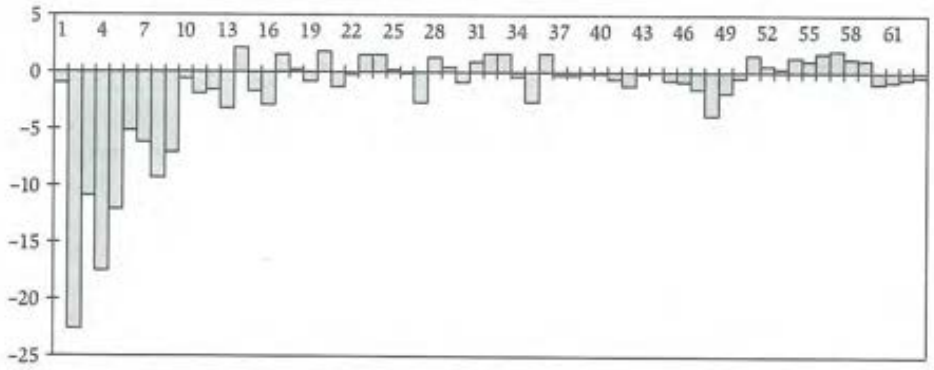


Zigzag scanning

(a)



(b)



(c)

Figure 3.9 DCT spectral block transmission order. (a) Zigzag scanning of the amplitude blocks. (b) Frequency coefficients scanned in zigzag order including dc component. (c) Frequency coefficients scanned in zigzag order without dc component.

Entropy Coding

The run length information occurs at a predetermined set of probabilities. Because of this, it is possible to employ entropy coding. As in Morse code, the most frequently occurring patterns are represented with the shortest codes, and the less frequently occurring patterns are assigned longer codes. This further reduces the number of bits needed to transmit the information. Huffman developed a method of creating such compression codes, called *Huffman codes*, when the probability of the message symbols is known. If another language is to be transmitted, another symbol table is used. If digitized video is to be transmitted, still another symbol table is applied.[17]

3.3.3 Coding of Moving Images

Now that we have completed the discussion of the compression of a still image (*intraframe coding*), we can extend the model to deal with moving images (*interframe coding*). Still-image compression ratios of 8:1 can typically be achieved with these techniques, but we still have a long way to go to reach a 50:1 ratio. Fortunately, all of the same principles we applied to a still image (transform coding, quantization, run length encoding, Huffman coding) can also be applied to the differences between one frame and the next. Two techniques are often used in conjunction in intraframe coding:

1. Calculate the difference between a macroblock in one frame and another (*picture differencing*).
2. Attempt to find the best match between a macroblock in one frame and that macroblock in another frame and transmit the motion vector (*motion compensation*).

Picture differencing breaks down if there is significant motion in the image. If you imagine a typical scene where the camera is panned across a scene, there are unlikely to be many similarities between a macroblock in one frame and that macroblock in the subsequent frame. The same is true for moving objects within the scene.

Motion compensation attempts to find where each macroblock has "moved" between frames, but it is unlikely that an exact match will be found, due to distortion caused by perspective and lighting changes. So when the best match is found, picture differencing is used to send the changes in the macroblock along with its new position.

Both of these techniques are used in conjunction in efficient video compression algorithms (such as MPEG).

Picture Differencing

In successive frames of video, each frame is typically very similar to the previous frame. If we take the limiting case of video of a still image, say, a landscape with

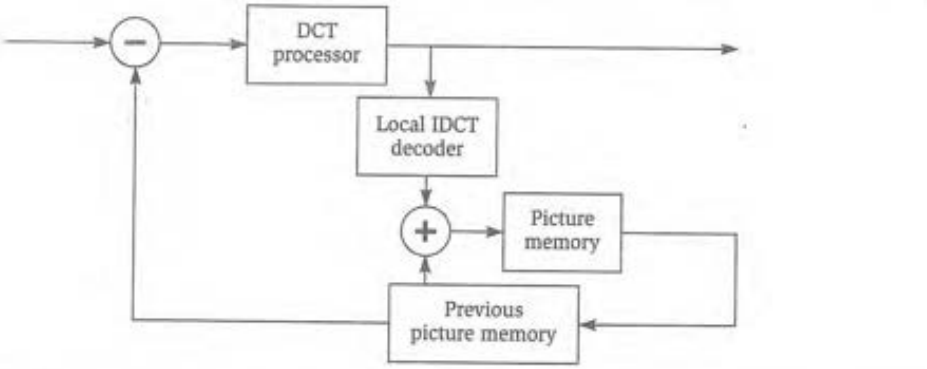


Figure 3.10 Interframe encoder.

no activity, there are no differences between frames. Sending one frame tells all. Succeeding frames simply can be repeated from the receiver's memory.

Figure 3.10 shows how the encoder can avoid sending the same frame repeatedly. When the equipment is turned on, all picture memories are empty. The first frame is sampled, and the DCT processor creates the frequency coefficient matrices for the luminance and color blocks of the picture and sends them to the IDCT decoder at the receiver site and to a local IDCT decoder. The local decoder produces the same image that will exist at the receiver. This image is then compared to the contents of the previous picture memory. Since there was no previous picture, the previous picture memory is empty, and there is nothing to add or subtract. The entire decoded picture is loaded into the picture memory. When the second frame is sampled, the previous picture is subtracted from it. Since they are identical, the subtractor produces zero values, the resulting coefficient matrix is all zeros, and the previous picture is unchanged. The process repeats itself, with no differences being transmitted to the receiver.

Figure 3.11 displays the decoder. Initially, its picture memories are empty. When the first frame's coefficient matrix arrives, it is decoded in the IDCT unit.

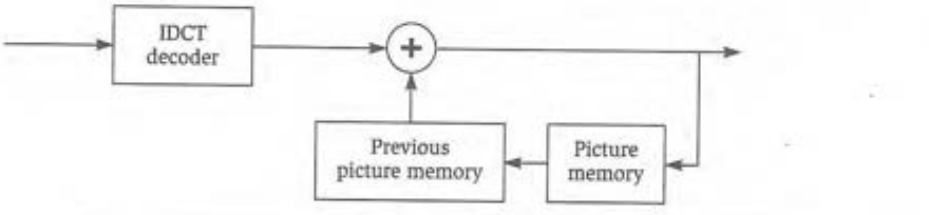


Figure 3.11 Interframe decoder.

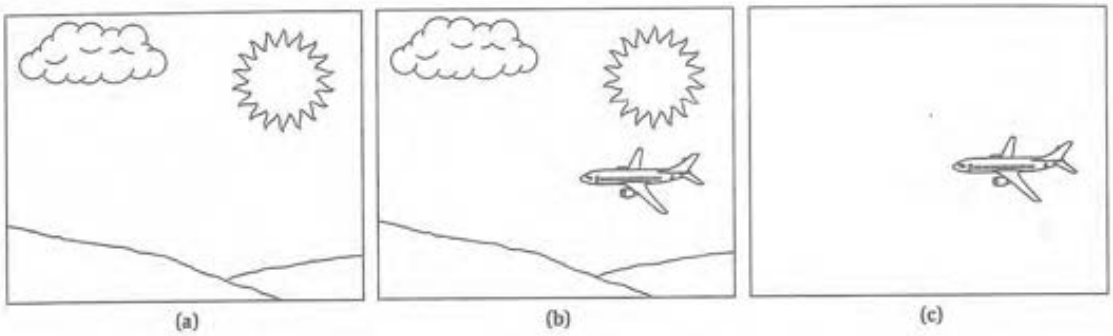


Figure 3.12 Differential transmission. (a) Original image. (b) Modified image. (c) Difference.

This is presented for display and for storage in the picture memory. At the time of the next frame, a zero coefficient matrix arrives. Nothing is added to the previous picture memory, and its contents are presented for display and are reloaded into the picture memory.

Watching a still picture on television, for example, the landscape image, shown in Figure 3.12(a), is not very interesting. Now let's assume that the picture changes because an airplane enters the picture (see Figure 3.12(b)). After the encoder subtracts the previous picture from the new picture, only macroblocks containing the image of the airplane have changed (see Figure 3.12(c)). These macroblocks may be encoded as differences from the previous picture, and a new coefficient matrix is transmitted to the receive site and the local decoder. (Note that the encoder may choose to intracode the macroblocks.) At the decoder, the coefficient matrix of the airplane is received and added to the previous landscape stored in the previous picture memory. It is presented for display and stored in the picture memory.

Two kinds of coefficient matrices have been transmitted in our example. The first coefficient matrix was a complete picture, called an intraframe (*I frame*), and all macroblocks are intracoded, or *nonpredicted*. The next frame contained information used to predict the contents of the picture memory for display; it is called a *P frame*. (Note that a P frame actually contains a mix of predicted macroblocks and nonpredicted macroblocks.)

In a perfect world, we'd need just one I frame and all the rest could be P frames. There are a couple of reasons why this is insufficient. First, if the viewer missed the I frame while "surfing" channels, he or she would not get a correct image. In our example, the viewer might get just the airplane and not the landscape! A second problem is that there are imperfections and distortions that accumulate. It is important to reset the process every so often to limit the propagation of errors. So I frames are sent several times a second to allow channel surfers to have a starting point and to limit the propagation of errors.

Motion Compensation

Motion compensation can be employed to significantly reduce the redundancy even further. For example, in Figure 3.13(a) and (b), we note that the image of the airplane is the same in all appearances. It is redundant to have to retransmit the airplane; once should be sufficient! It should be adequate to simply convey an instruction to move the airplane to the new position. Of course, it is not that simple.

Motion compensation is based on the observation that much of the motion in images consists of a displacement of an element of the image. Motion compensation takes each macroblock and compares it with all possible macroblocks formed by making horizontal and/or vertical displacements in small increments (typically half-pixel). When the best fit is found, instructions are conveyed to the receiver to work with the contents of the macroblock and, when done, to relocate them to their new position with a *motion vector*. In our example, the encoder motion compensation would determine that the airplane has simply been moved horizontally. Instructions to the receiver would tell it to take the contents of a number of blocks in memory and move them to other locations that are some number of pixels displaced horizontally and vertically.

The motion compensation algorithm will usually find that the macroblock match is not perfect, and so it encodes any changes and transmits them with the motion vector. In our example, if the airplane made a slight rotation, simply moving it horizontally and vertically would not be enough. The difference between its previous pixels and its new pixels would be encoded. In most cases, this would be just a small fraction of the total pixels. Most would be unchanged. The resulting instructions to the receiver include taking blocks from memory, modifying them slightly, and then relocating them. When they are relocated,

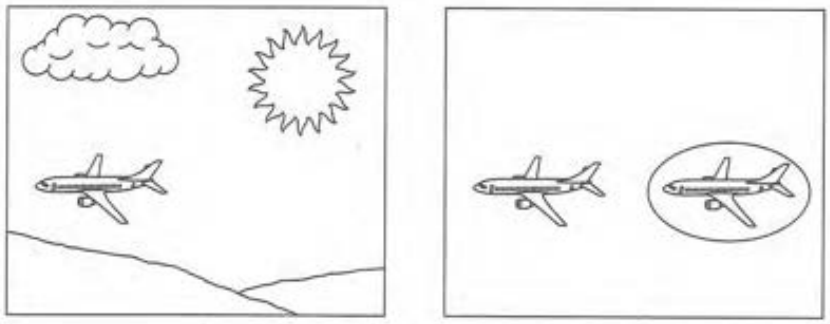
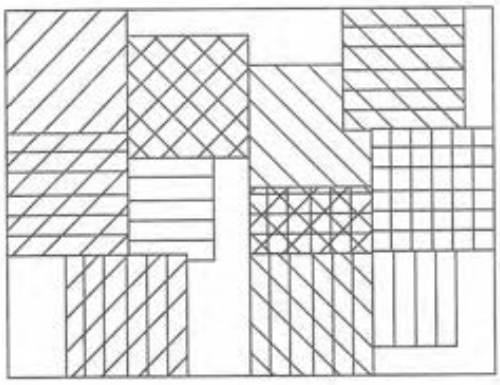


Figure 3.13 More complex motion. (a) Further modified image. (b) Differences.

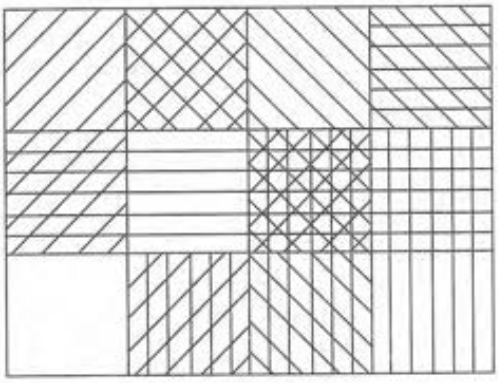
parts will overlay other blocks, and parts will be covered by other blocks. That information must be conveyed as well. Holes in the image will be created. They must be filled. The various parts of the image may move in different directions by different amounts. The motion compensation process can be very sophisticated to aggressively reduce the amount of information that needs to be transmitted (see Figure 3.14). Areas in the preceding frame of Figure 3.14 are found that best match the areas in the current frame. These are transmitted along with instructions for their new position and the differences needed to update them. This minimizes the amount of information that must be conveyed to the receiver.

Motion compensation makes the encoder much more complex than the decoder, which is acceptable because there are many more decoding



Preceding frame

Areas in preceding frame that most closely match corresponding areas in current frame



Current frame

Figure 3.14 Motion compensation.

sites than encoding sites. Since motion compensation is done at the encoder, viewers will enjoy improved results with their existing decoders as encoder motion compensation technology advances. The goal is to make receivers inexpensive.

B Frames and Groups of Pictures

Consider the airplane and landscape example shown in Figure 3.15. Five locations of the airplane occur. When the airplane moves from one position to another, it uncovers a piece of sky, which is present in subsequent images. If we had one of the subsequent images already in memory, for example, Figure 3.15(c), we could take the piece of sky needed from a "future" image. We could "predict backward" in time. The ability to predict both forward and backward in time is possible if we introduce a slight delay in the transmission and add memory to the receiver for more than one frame of video. The frames that allow bidirectional prediction are called *B frames*. Although they add cost to the decoder, they increase transmission efficiency. The trade-off in cost and complexity is governed by the cost of the memory. The cost of memory has come down substantially, so B frames are a worthwhile investment because bandwidth is the ultimate scarce resource.

We have defined three kinds of pictures: I frames, P frames (or predicted frames), and B frames (or bidirectionally predicted frames)*. They are clustered into a *group of pictures* (GOP) (see Figure 3.16). In Figure 3.16(a), the frames are presented in the order in which they are displayed. In order to minimize receiver complexity, they need to be transmitted in the order shown in Figure 3.16(b). In a sophisticated encoder, the GOP flow is interrupted at a scene change, and an I frame is transmitted (recall that an I frame contains all of the information needed to decode the frame within the frame itself and thus is not dependent on adjacent frames). However, I frames require much more data than other frames (a ratio of 5:3:1 for I:P:B frame size is typical). Therefore, the flow of data is very uneven, and the decoder must have a buffer of sufficient size to even out the flow.

Field-Based Prediction

All of the discussion so far has dealt with video as a sequence of frames. If the source material is generated by interlace scanning (as is the case with most video material), each frame is actually composed of two fields separated in time by the field period. For scenes with relatively little motion, the "double image" is

* More correctly, but less commonly, these are known as *I pictures*, *P pictures*, and *B pictures*. The *frame* terminology probably got started because MPEG was originally designed to code purely frame-based material and seems to have stuck.

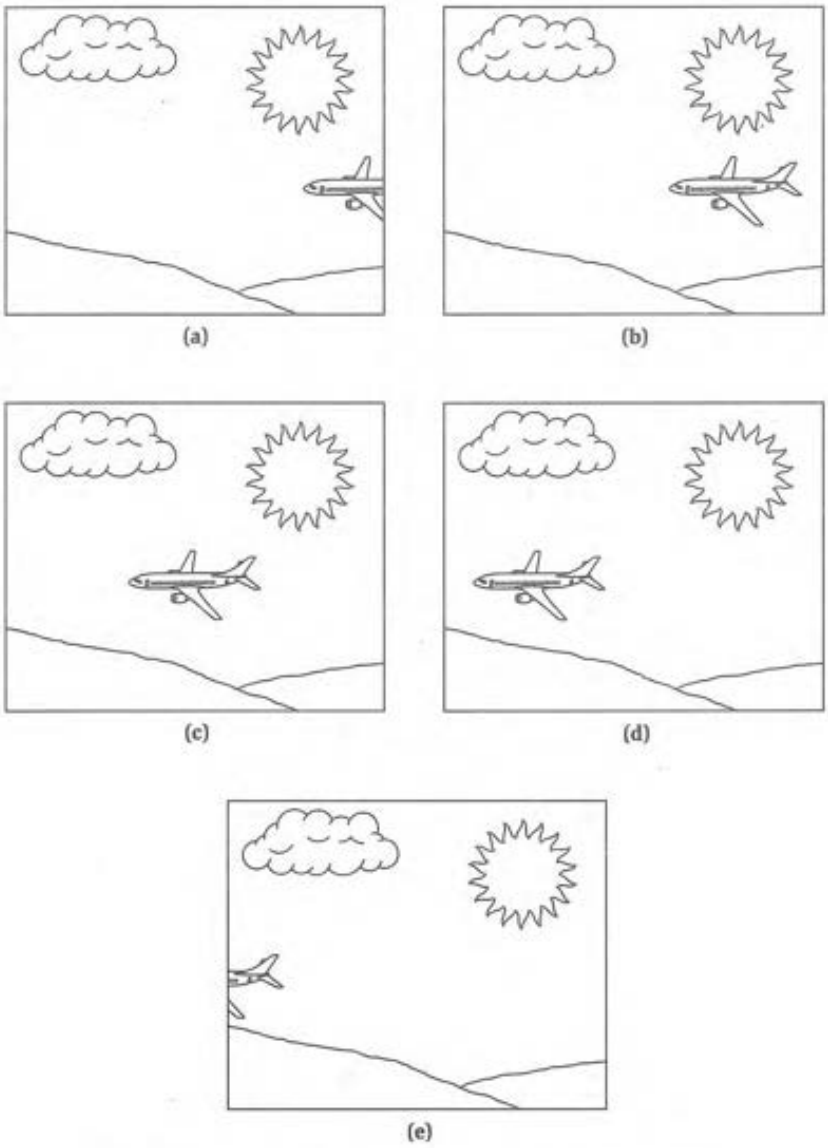


Figure 3.15 Motion compensation example.

not a problem and pure frame-based encoding may produce acceptable results. However, when there is rapid motion, the motion compensation algorithm needs to adaptively move into its field-based mode. The decision to code the field or the frame can be made at the macroblock level.

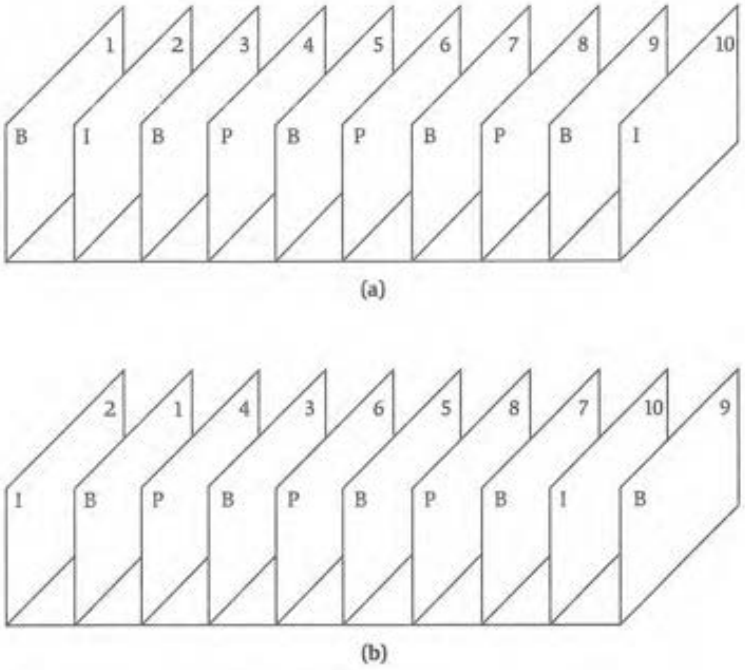


Figure 3.16 Order of transmission and display of I, P, and B frames. (a) Frame type and display order. (b) Frame type and transmission order.

If the field is coded, the values for the block are taken from alternate lines in the frame buffer — essentially reconstructing the field. After DCT, the coefficients for field 1 are sent, followed by those for field 2 for that macroblock. Picture differences and motion compensation for the block can reference either of the two fields in the preceding or succeeding frames.

Obviously, adaptive field/frame-based coding represents a considerable increase in complexity at the encoder and decoder, but it is essential to provide good results for interlaced material at moderate bit rates.

3.3.4 Digital Artifacts

Since each block of the image is treated separately, it is easy to appreciate that there may be differences in decoded video at the boundaries of the blocks. Techniques exist for smoothing over these boundaries. A digitally compressed image with inadequate data will occasionally display the boundaries of the blocks. This is called *blocking noise*.[18]

Recall that the 8-by-8 DCT and IDCT perform lossless conversion between the video and frequency space representations to the limits of computational precision. However, once the coefficients are quantized, the lossless nature of the computations is abandoned. Some distortion is introduced. If the distortion is minor, it might not be noticeable. But even if it is noticeable, it might not be objectionable; it may just blend in and get lost. However, under some circumstances, sharp edges will yield attenuated ripples. This is called the *Gibbs phenomenon*, after the mathematician who studied it.

If a sharp point exists in the image, the Gibbs phenomenon will exist in two dimensions. The resulting image distortion has been called *mosquito noise* or *the attack of the killer bees*. This artifact appears as a cluster of mosquitoes or killer bees surrounding the sharp point in the image. If a sharp edge exists in the image, ripples may precede and follow it.

3.3.5 Information Carried in the Vertical Blanking Interval

The astute reader will have noticed that only the active video lines in the NTSC signal are encoded. Clearly, there are several information fields normally carried in the vertical blanking interval (VBI), which would not make it to the television receiver:

1. Closed captioning, carried in line 21, field 1[19]
2. Nielsen source identification (SID) automated measurement of lineups (AMOL)
3. Teletext
4. VBI time code (VITC)
5. Vertical interval test signals (VITS)

Two SCTE standards[20,21] have been developed to encode these lines and insert them into the digital bitstream for carriage to the STT. Since the VBI information is considered inseparable from the video itself, it is carried in the video elementary stream by the picture user data field specifically designated for this purpose (see MPEG-2 Transport discussion later). After the video is decoded, the picture user data is used to reconstruct the VBI data, and it is inserted into the NTSC output signal by the NTSC encoder. In this way, backward compatibility is maintained with the installed base of analog televisions, which expect to see closed caption information in the VBI of the baseband analog signal.

Digital television has its own standard for closed captioning, specified by EIA-708,[22] which includes additional features above and beyond EIA-608, such as popup windows, extended characters sets, and the like. The FCC imposed a deadline of July 2002 to include EIA-708-compliant closed captions in digital television transmissions.

3.3.6 High-Definition Television

The discussion of digital video compression so far has focused on standard-definition television signals. The principles apply equally well to high-definition television, except that picture resolution and bit rates are increased considerably.

Sampling

HDTV has two main variants — a 720-line progressively scanned system (720p) and a 1080-line interlace scanned system (1080i). The latter is based on the SMPTE 240M standard for 1125/60-line HDTV production systems.[23] Sampling parameters for the two systems is shown in Table 3.2 (standard-definition television is included for comparison).

A 1080i picture has six times as many active pixels as a standard-definition picture, and hence the raw information rate is increased by a factor of 6. The development of high-definition encoders has provided a significant challenge to chip manufacturers, a challenge that has been met by a combination of parallel processing and higher clock rates, made possible by the reduction in integrated circuit line widths.

Block formation

Although the picture resolution is higher, the block size remains the same; there are just more blocks per picture. In addition, macroblocks are formed in the same way. Standard-definition television uses the MPEG-2 main profile at main level (MP@ML), while high-definition television uses the MPEG-2 main profile at high level (MP@HL), (see Section 3.3.7). The SCTE has defined the allowed set of resolutions and frame rates that are supported by cable systems as shown in Table 3.3.

3.3.7 MPEG Standards

The Motion Picture Experts Group (MPEG) of the Geneva-based International Standards Organization (ISO) has been active in video compression standards since the early 1990s. All MPEG standards are based on a set of principles that have allowed tremendous innovation while maintaining an interoperable standard.

Table 3.2 Standardized Video Input Formats (from [26], used with permission)

Video standard	*Active lines*	*Active samples/line*
ITU-R BT.601-5	483	720
SMPTE 296M[24]	720	1,280
SMPTE 274M[25]	1,080	1,920

Table 3.3 Compression Format Constraints for Digital Video (from SCTE 43 2001[26] used with permission)

Vertical resolution	*Horizontal resolution*	*Aspect ratio*[a]	*Frame rate*[b]	*Progressive sequence*[c]
1,080	1,920	1, 3	1, 2, 4, 5	1
1,080	1,920	1, 3	4, 5	0
1,080	1,440	3	1, 2, 4, 5	1
1,080	1,440	3	4, 5	0
720	1,280	1, 3	1, 2, 4, 5, 7, 8	1
480	720	2, 3	1, 2, 4, 5, 7, 8	1
480	720	2, 3	4, 5	0
480	704	2, 3	1, 2, 4, 5, 7, 8	1
480	704	2, 3	4, 5	0
480	640	1, 2	1, 2, 4, 5, 7, 8	1
480	640	1, 2	4, 5	0
480	544	2	1	1
480	544	2	4	0
480	528	2	1	1
480	528	2	4	0
480	352	2	1	1
480	352	2	4	0

MPEG-2 coded values:

[a] 1 = square pixels; 2 = 4:3; 3 = 16:9.

[b] 1 = 23.976 Hz; 2 = 24 Hz; 3 = 25 Hz; 4 = 29.97 Hz; 5 = 30 Hz; 6 = 50 Hz; 7 = 59.94 Hz; 8 = 60 Hz.

[c] 0 = interlaced scan; 1 = progressive scan.

- MPEG standards do not specify the operation of the encoder, only that of a reference decoder and the syntax of the bitstream that it expects. This allows encoder implementations to improve over time, for example, as better motion compensation search algorithms are developed.
- MPEG standards are a toolbox approach — various features are added from one "profile" to another, each higher profile adding complexity and efficiency.

MPEG-1

MPEG-1,[27] which was standardized in November 1992, introduced efficient coding for frame-based (progressively scanned) material based on DCT coding of 8 by 8 blocks. MPEG-1 is based on the intraframe coding principles from JPEG,

but it adds interframe coding techniques, including motion compensation and the associated difference pictures, the so-called P and B frames. MPEG-1 supports only low resolutions (352 × 240), which is insufficient for entertainment-quality video. (However, it was quickly extended to higher resolutions, for example, the Full Service Network[28] used so-called "MPEG-1.5" at a resolution of 352 × 480. DirecTV also launched with MPEG-1 at higher resolution, subsequently migrating to MPEG-2 before the transport packet structure was finalized.) Although MPEG-1 works well for progressively scanned material (for example, film), it requires high bit rates (6 Mbps or greater) to compress NTSC video. This is primarily because MPEG-1 does not support field-based coding.

MPEG-2

MPEG-2 was standardized in November 1994. MPEG-2 adds support for adaptive field/frame coding, which is particularly for encoding interlaced material with high motion content (for example, sports). (MPEG-3 was originally targeted at HDTV, but it was abandoned when it became apparent that MPEG-2 was easily extendable to support HDTV.)

MPEG-2 has been a tremendously successful standard. It has been adopted by the television and cable industries worldwide and also by DVD. MPEG-2 provides a toolbox comprising a collection of image-processing components; a specific application uses appropriate portions of the standard. Table 3.4 summarizes the profiles and levels defined by the MPEG-2 standard (some levels and profiles have been omitted for clarity). The "level" describes the amount of video resolution. The "profile" describes the principal features of that collection of components. The general-purpose profile is called the *main* profile. The *simple* profile drops the B frames and simplifies the decoder by reducing the

Table 3.4 MPEG-2 Levels and Profiles (Subset)

		Profile			
Level	*Pictures Sampling*	*Simple I, P 4:2:0*	*Main I, P, B 4:2:0*	*High I, P, B 4:2:0 or 4:2:2*	*4:2:2 (studio) I, P, B 4:2:2*
High	Resolution		1920 × 1152	1920 × 1152	
	Frame rate		60 fps	60 fps	
Main	Resolution	720 × 480	720 × 480	720 × 480	720 × 512
	Frame rate	29.97 fps	29.97 fps	29.97 fps	29.97 fps
	Resolution	720 × 576	720 × 576	720 × 576	720 × 608
	Frame rate	25 fps	25 fps	25 fps	25 fps
Low	Resolution		352 × 288		
	Frame rate		30 fps		

memory and processing required. The trade-off is that the bit rate generally goes up. Standard-definition television uses the main profile at main level (MP@ML), while HDTV uses the main profile at *high* level (MP@HL), although there is a separate high profile designed with HDTV in mind. In addition, a 4:2:2 profile is defined for studio use.

MPEG-4

MPEG-4 was originally focused on "very-low-bit-rate coding" (as described in the 1993 RFI), but was changed the following year to "coding of audiovisual objects." MPEG-4 is very different in focus from previous MPEG standards, in that it focuses on multimedia and interactivity. Visual scenes are broken down into objects and sent as separate layers, which are composed at the decoder.

MPEG-4 also introduces some innovative techniques for very-low-bit-rate coding. These include mapping of images onto a computer-generated mesh — it is possible to map facial expressions and send them independently from a person's image and then reconstruct the animated face at the receiver.

Most of MPEG-4 is not intended for entertainment-quality video and is not very relevant to the cable industry. However, MPEG-4 Part 10, Advanced Video Coding, is a new standard for entertainment-quality video that offers significantly lower bit rates than MPEG-2.

MPEG-4 Part 10, Advanced Video Coding (AVC)

MPEG-4 Part 10 was completed in March 2003. It was developed by the Joint Video Team (JVT), a joint effort of the ITU-T Video Coding Experts Group (VCEG) and MPEG. It is also known as H.26L.

MPEG-4 AVC adds improvements to DC and AC prediction, sprite coding, error resiliency coding, nonlinear dc quantization, and has more accurate motion compensation (quarter-pel over MPEG-2's half-pel). MPEG-4 AVC also uses spatial prediction for intraframe coding to reduce the size of I frames. The basic spatial transform is an integer transform (an approximation of the DCT) that operates on a smaller, 4×4 pixel block. To improve motion compensation efficiency, MPEG-4 AVC adaptively uses different block sizes (4×4, 8×4, 4×8, 8×8, 16×8, 8×16, or 16×16) and multiple reference frames. Finally, postfiltering of the image is much more sophisticated, to help reduce blocking artifacts, and includes a deblocking filter in the prediction loop.

Just how much better is MPEG-4 AVC than MPEG-2? It is still too early to say at the time of this writing, but substantial improvements in coding efficiency are already being demonstrated by MPEG-4 AVC implementations. As with MPEG-2, we expect to see an improvement in coding efficiency over time. For example, from the MPEG-2 curve shown in Figure 3.17, it is apparent that MPEG-2 has almost reached the limitations of the fundamental principles upon which it is based. Bit rates much below 2 Mbps for entertainment-quality video do not seem

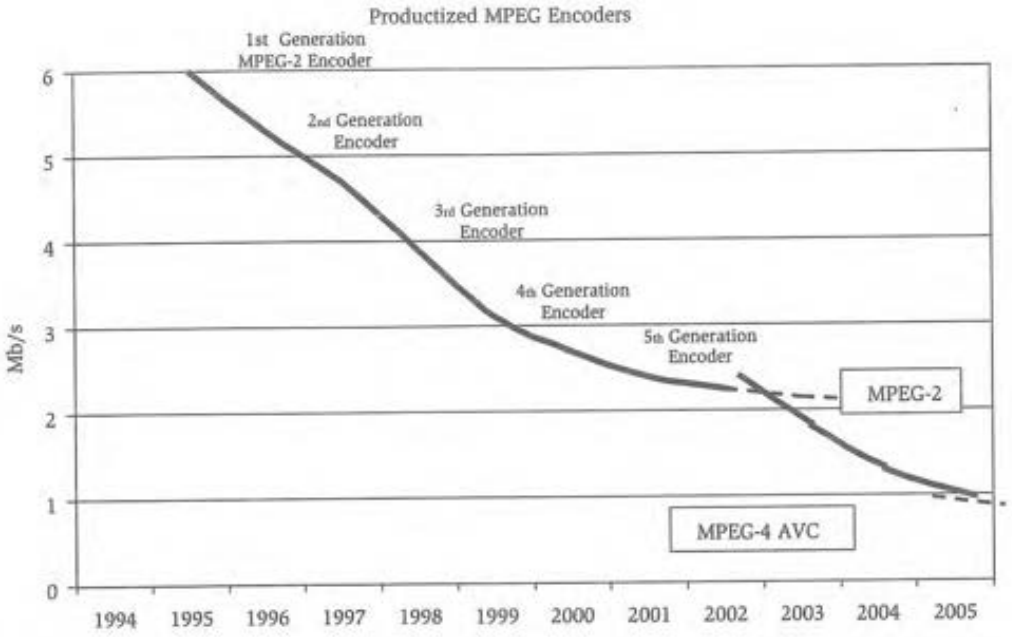


Figure 3.17 MPEG-4 AVC versus MPEG-2 bit-rate. (Courtesy of Harmonic Inc.)

likely. In contrast, MPEG-4 AVC shows the potential to be at least twice as efficient as MPEG-2 for the same perceptual quality.

However, there are significant issues of backward compatibility that prevent cable operators from adopting of MPEG-4 AVC, attractive as it may be from the point of view of bandwidth efficiency. SCTE digital television standards are based on MPEG-2 video compression. Moreover, a migration to MPEG-4 AVC would necessitate the wholesale replacement of tens of millions of decoders. Since video on demand is not currently subject to SCTE standards, it could employ MPEG-4 AVC to achieve more VOD streams per QAM channel, but at the cost of a new STT per VOD household.

MPEG-4 AVC is being seriously considered for high-definition DVD9 by the DVD forum. DVD9 is a single-sided, dual-layer DVD format with a total capacity of 8.5 GB. Toshiba demonstrated storage of approximately 135 minutes of high-definition video using MPEG-4 AVC encoding at 7 Mbps at the 2003 Consumer Electronics Show.

The issues of backward compatibility do not affect MPEG-4 AVC adoption in the Internet environment. Several software decoders were already available at the time of writing, and Apple has adopted MPEG-4 AVC as the core for QuickTime. Internet streaming media will probably enjoy a considerable advantage in

coding efficiency over broadcast television services (cable, satellite, and terrestrial broadcast).

3.3.8 Other Approaches to Video Compression

A variety of other approaches to compression exist besides the discrete cosine transform. The DCT has enjoyed the most attention because it is based on the Fourier transform, which has been studied more than any other transform. The Fourier transform has other broad applicability to the study of spectra and signal analysis. It has received extensive funding in research on radar, sonar, and deep-space signal analysis. Efficient computational methods have been developed to apply it. Custom-designed integrated circuits have been constructed to implement very fast conversion of waveforms to spectra, and vice versa. The DCT is a natural starting point for video compression.

Other approaches will continue to be studied and may at some point close the gap that currently exists between DCT-based and other compression techniques. Efficient software and hardware may develop for these other approaches, and eventually the DCT may be displaced. We will summarize three alternatives — wavelets, fractals, and vector quantization — that have all been around for a considerable time but have not seriously challenged DCT-based approaches.

There will continue to be further advances in video compression (like MPEG-4 AVC). It is therefore important to avoid building regulatory or other mechanisms that will stand in the way of implementing advancing technology. The recent history of the personal computer industry has shown that the marketplace will tolerate some degree of difficulty with interoperability and compatibility if the technical advances are substantial. This comes at a heavy toll in consumer frustration and sometimes anger. However, the old model of rigid compatibility may fall to the rapid advance of technology. In the end, the marketplace decides.

Wavelet

Wavelet transforms appear to be more "scalable" than other techniques. That is, they have the ability to serve a low-resolution display and then, with little waste of what has been transmitted, add more and more detail. A low-resolution display is able to use simpler circuits to extract what is needed, whereas a higher-resolution display can support the cost of more complex circuits and extract more information from the signal. A special class of filters is used to implement wavelets. As more research is done, more cost-effective implementations will become available.

Fractal

Briefly, fractals are equations that describe images. They may be the most efficient compression method, representing the most complex images with the least data. The challenge with fractals has been in finding the equations.

Vector Quantization

Vector quantization involves approximating the image out of a prestored table of image elements (sometimes called *stamps*). This "library" is used to build up the image. Choosing the elements to place in the library is critical to the success of vector quantization. Simpler images place lower demands on the library; images that are more complex are more challenging.

3.4 Digital Audio Compression

The human auditory system is considerably less tolerant of imperfections than the human visual system. Compression of quality audio is much more difficult than the compression of video. Compression ratios of only 4:1 or 8:1 are achieved in audio, whereas video enjoys compression ratios of 30:1 or 50:1. Fortunately, the audio bandwidths are small, and the resulting data rates are not too burdensome.[29]

Two major competitive methods of audio compression are in widespread use in digital television systems: MPEG-1 Layer 2 and Dolby AC-3. Much of Europe and the U.S. DSS satellite system have chosen MPEG-1 Layer 2, whereas the U.S. cable industry and the U.S. broadcast standards use AC-3. DVD uses Dolby AC-3 for 60-Hz systems (mostly U.S. and Japanese). Some DVDs (digital video discs) use MPEG-2 audio, but only the backward-compatible mode to MPEG-1 Layer 2, for 50-Hz scanning systems (mostly European).

A third method, MPEG-1 Layer 3, has become extremely popular for audio-only applications, and is generally known by its Windows file-extension name, "MP3."

3.4.1 Principles of Audio Compression

As with video, compression is achieved by the quantization of the frequency components of the signal.

The ear is extraordinarily sensitive to certain impairments in the audio signal, and spectral flatness is very important. Nevertheless, the ear is most sensitive to frequencies between 2 kHz and 5 kHz, and sensitivity drops quite rapidly above and below these frequencies. In fact, at 50 Hz and 17 kHz the sensitivity of the ear is about 50 dB lower than at 3 kHz.

Another effect of the psychoacoustic system is audio masking. A louder sound will mask a quieter one, especially if they are close in frequency. (This effect causes the white noise of the air conditioning fan in a hotel room to mask sounds from outside the room as long as they are not too loud.) The phenomenon is quite complex and very nonlinear. It is more effective when the quieter sound to be suppressed is higher in frequency than the louder sound. The extent to which a louder sound masks a quieter sound is called the *psychoacoustic masking threshold*. Perceptual coding takes advantage of this masking effect.

Figure 3.18[30] shows a high-level block diagram of an audio encoder–decoder pair. In the encoder, an analysis filter bank divides the input signal into narrow frequency bands, each of which may then be quantized and coded independently. The perceptual model adjusts the quantization level of each band such that quantization noise is ideally just below the masking threshold. The resultant bitstream is encoded and transmitted over a data link to the decoder. At the decoder, the inverse functions are applied to reconstruct the audio signal for playback to the listener.

Here are the main steps in the audio compression process.

1. Sampling and preprocessing
2. Sub-band coding
3. Transform coding (optional)
4. Quantization according to psychoacoustic masking threshold
5. Multichannel bit allocation (optional)
6. Huffman coding

Entertainment-quality audio is generally sampled at 48 kHz, with quantization to 16, 18, 20, or 24 bits per sample. (Compact discs are the exception, with 44.1 kHz-sampling at 16-bit quantization.)

Sub-band coding divides the audio spectrum into narrow frequency bands and codes each sub-band separately. The advantage of this approach is the quantization of each sub-band can be set independently according to (a) the sensitivity of the ear at that frequency and (b) the masking effect of signals in adjacent sub-bands. More advanced audio coding systems may use transform coding in conjunction with sub-band coding or to completely replace sub-band coding.

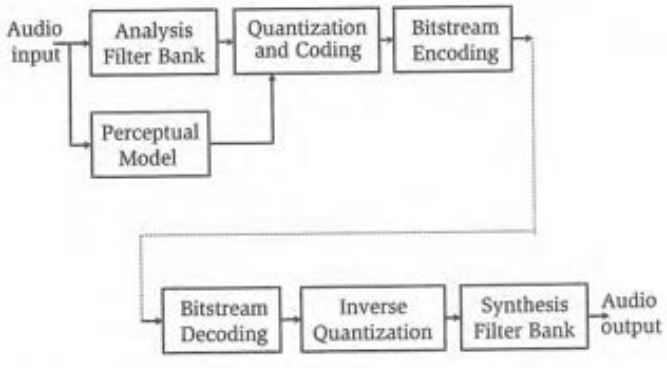


Figure 3.18 Audio Encoding and Decoding.

As each channel is coded, bits are dynamically allocated between sub-bands to achieve the best fidelity for a given bit rate. Huffman coding is used to code frequently quantized values in the same way as for DCT coefficients in video coding.

For multichannel audio, there is often redundancy between channels that can be exploited by multichannel bit allocation schemes. In particular, the ear locates sound sources better at some frequencies; outside of these frequency ranges two channels can be "bridged" together.

There are many excellent texts that discuss audio compression in more detail.[31,32]

3.4.2 AC-3 (Dolby Digital)

The AC-3 system has 5.1 channels of audio: front left, front right, center, surround left, surround right, and low-frequency enhancement (LFE). The bandwidth of the channels is around 20 kHz, with the exception of the LFE channel, which is limited to 120 Hz. This gives rise to the ".1" of the 5.1 channels. Since low-frequency sounds are not localized, only one speaker is needed for the low frequencies, and it can carry the low frequencies of all the channels. Additionally, the LFE speaker (popularly known as a "subwoofer") can be placed almost anywhere in the room.

Audio is compressed by first converting it to digital form with a sampling rate of 48 kHz. The audio is then segmented into *frames* of samples 1,536 samples long. The analog audio is quantized to at least 16-bit resolution. The frame is divided into six 512-sample *blocks*. Since the human auditory system is so sensitive to imperfections, half of the samples in each block overlap half of the samples from the adjacent blocks. That is, there is a 25% overlap with the preceding block and a 25% overlap with the following block. The sensitivity of the human auditory system to discontinuities requires doubling the number of transmitted samples! The samples are *windowed*; that is, the set of samples is multiplied by a weighting function that tapers the amplitude of the samples at the beginning and at the end of the sample set. The receiver blends the ending tail from one set with the beginning tail from the next set to avoid sudden changes or clicks or pops.

The audio samples are transformed into the frequency domain as a set of spectral data points using a modified version of the discrete cosine transform called the *time division aliasing cancellation* (TDAC) transform. Its special properties allow recovery from the need to have block overlap that would otherwise generate twice as many frequency coefficients. The result is that the 512 samples of an audio block yield only half as many, 256, frequency coefficients. These frequency coefficients divide the audio spectrum into 256 subbands of less than 100 Hz each. Up to this point, the processing of the digital samples is fully reversible, and only digitization has been accomplished. Compression has not yet been achieved.

Data compression is achieved by quantization and entropy coding. The psychoacoustic model dictates the degree of coarseness that can be accepted in the quantization of the audio spectrum data. The frequency coefficients are transmitted in the mathematical format known as *scientific notation*. A basic number is multiplied by an exponent. The basic number is called the *mantissa*. The same exponent generally applies to several adjacent blocks of samples, and data transmission can be reduced by taking advantage of this. Only changes are conveyed. Extensive techniques exist for reducing the amount of data that must be transmitted. The optimum bit rate for 5.1-channel Dolby AC-3 is 384 kb/s, although other rates are possible, yielding varying qualities of the resulting decoded audio signals.

AC-3 provides two important features to manage audio level and dynamic range at playback.

Dialog normalization: A subjective loudness of the dialog is explicitly coded in the AC-3 bitstream to allow balance to be achieved from one program segment to another. This is necessary because the level of the dialog may vary greatly between, for example, a movie and a commercial.

Dynamic range description: The dynamic range of the playback may also be adjusted at playback. This allows a listener with a home theater to enjoy the full dynamic range of a movie while also allowing the same movie to be enjoyed on a television with a small speaker.

Finally, AC-3 also defines an audio frame structure, which includes a method for synchronization and CRC-16 error detection codes. If an error is detected in the bitstream, the decoder will mute the audio and attempt to resynchronize. AC-3 was selected as the audio standard by the Advanced Television Systems Committee (ATSC) Digital Television Standard.[33] It is also specified by the standards for digital cable systems of the Society of Telecommunications Engineers (SCTE).

3.4.3 MPEG Audio

MPEG-1 Audio

MPEG-1 audio[34] coding defines three *layers*, which are the audio equivalent of video profiles. Each layer adds complexity but reduces bit rate. MPEG-1 audio was originally based on principles from the MUSICAM (masking-pattern adapted universal sub-band integrated coding and multiplexing) system.

MPEG-1 layer 2 is commonly used in satellite applications and in European broadcast applications. It uses 32 sub-bands (using quadrature mirror filtering) and codes entertainment-quality music at 128 Kbps per channel.

MPEG-1 layer 3 further subdivides the 32 sub-bands of layer 2 into 576 sub-bands by using a frequency transform (modified discrete cosine transform). It can provide about the same quality as MPEG-1 layer 2 using only 64 Kbps per

channel. MPEG-1 layer 3 is better known as "MP3" (from the file extension .mp3) and has become the de facto audio encoding algorithm used for "ripping" CDs and sharing them on the Internet. MP3 is also widely used in portable music players.

MPEG-2 Audio

MPEG-2 audio has two main variants: backward-compatible audio[35] and advanced audio coding (AAC).[36] Backward-compatible audio adds support for 5.1 multichannel audio while maintaining bitstream compatibility with MPEG-1. AAC is a second-generation algorithm that provides a considerable increase in coding efficiency over MPEG-1 audio. Good results are obtained at 64 Kbps per channel. AAC is also very flexible, providing up to 48 channels of audio, 16 low-frequency effects channels, 16 overdub/multilingual channels, and 16 data streams.

3.5 Digital Audio–Video Transport

So far, the discussion has covered how we take individual video feeds and audio feeds and convert them into "elementary streams," which are their compressed digital bitstream equivalents. The next step is to combine these elementary streams into a *transport stream* with some kind of structure so that an STT can select and decode the digital bitstreams according to our familiar program paradigm for broadcast television.

Time division multiplexing (TDM) allows multiple lower-rate bitstreams to be combined to form a single higher-rate bitstream. However, familiar TDM methods used for telephony (for example, SONET) are designed to carry many channels that are all exactly the same rate. These would not be suitable for video and audio elementary streams with arbitrary, and possibly variable, transmission rates.

An approach more like asynchronous transfer mode (ATM) is required to allow each elementary stream to be transmitted at the precise rate it requires. ATM approaches divide the bitstream into fixed-length *packets* (also known as *cells*) and allocate packets between channels based on their required bit rate. A wider channel is created by allocating more packets per second than a narrower channel.

The MPEG-2 Systems Standard[37] defines an ATM-like method to transport digital video, audio, and auxiliary data in a multiple program transport stream (MPTS). It is the basis for the ATSC Digital Television Standard and is specified by the SCTE Standards for Digital Television.

3.5.1 MPEG-2 Systems Layer

The MPEG-2 systems layer defines the two different ways to multiplex elementary streams from audio and video encoders, as illustrated in Figure 3.19:

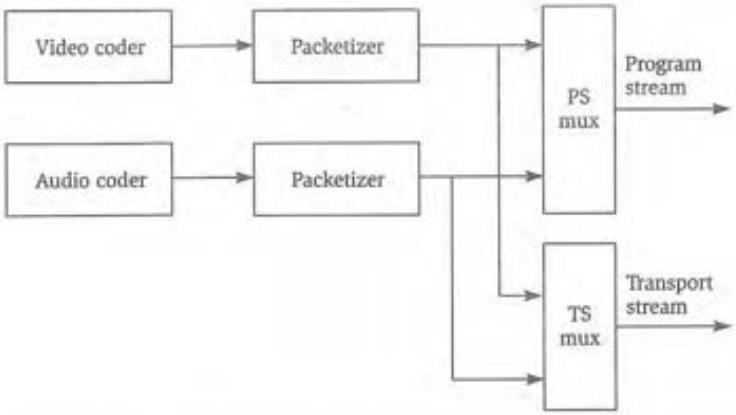


Figure 3.19 Coding, packetization, and multiplexing.

Program stream — designed for storage-based applications, for example DVD.[38] The program stream format is not generally used in cable television and will not be discussed further.

Transport stream — designed for transmission over a constant-delay network, for example, terrestrial broadcast and cable systems. The remainder of this discussion will focus on the transport stream.

Before the individual video or audio elementary bitstreams are combined to form a transport stream, they must have timing information inserted to describe the rate at which they should be displayed at the decoder, and they must be packetized so that they can be recovered at the receiver.

Figure 3.20 shows the structure of the packetized elementary stream (PES) packet. The start code prefix identifies the start of the packet. The stream ID identifies the type of stream (16 for video, 32 for audio). The packet is of variable length, as described by the PES packet length. A great deal of information is provided in the PES header (signaled by PES header flags), most importantly the presentation time stamp (PTS) and decode time stamp (DTS), which are used by

3 bytes	1 byte	2 bytes	2 bits	14 bits	1 byte	Variable length	
Packet start code prefix	Stream ID	PES packet length	10	PES header flags	PES header length	PES header fields	PES packet data block

Figure 3.20 Packetized Elementary Stream (PES) structure.

the receiver to synchronize the playout of audio and video. Since the PES packet is intended for error-free environments, there are no fields for error detection or protection.

The MPEG-2 systems transport stream supports an important set of features for broadcast digital television systems:

1. Multiplexing
2. Navigation
3. Timing distribution
4. Conditional access
5. Splicing

Multiplexing

Multiplexing allows many packetized elementary streams to be combined into a *multiple-program transport stream* (MPTS). This has numerous advantages for the operator. It allows the digital channel to be packed with multiple programs (analogous to television channels) so that the capacity of the channel is used efficiently. Additional auxiliary data can also be combined into the multiplex to carry program guide information, conditional access control messages (see Chapter 21), emergency alert information, and so on.

Figure 3.21 displays how packetized elementary streams are time division multiplexed to form a transport stream. Transport streams are identified by a unique packet identifier (PID) so that they can easily be filtered and routed to the different circuits and subsystems in the receiver that process and decode those data streams into the output signals that drive the presentation devices. Video and audio data must be intermixed in a manner that ensures that sufficient audio and video data arrive for their respective processors so that the sound and picture can be reproduced in synchronism. This figure is generic, and the concept applies to all transmission paths. The transport stream can be a

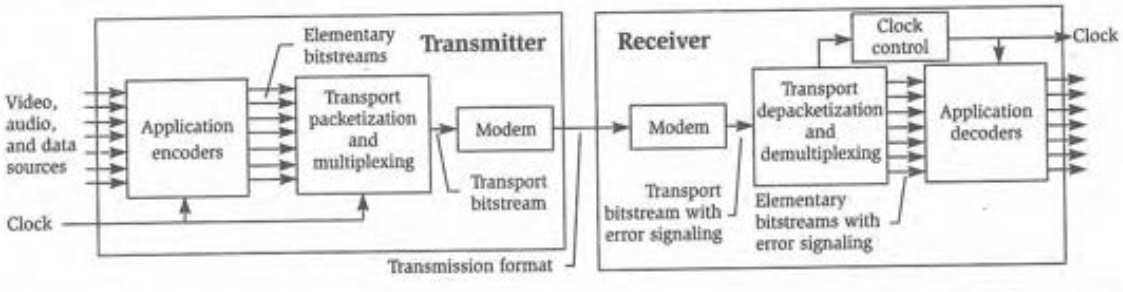


Figure 3.21 Transport packetization and multiplexing.

combination of a variety of packetized elementary streams. For example, the same video PES can be associated with several audio PESs in multiple languages. Likewise, the same audio PES can be associated with several video PESs, providing multiple views of the same athletic event. This packetization is necessary if there is to be local insertion of programming and commercials. This, of course, is a part of the current business structure in analog systems and is a critical part of the economics of the business.

Figure 3.22 shows the MPEG-2 transport packet format. Packets are fixed-length (188 bytes) with a fixed-length header (4 bytes). A sync byte having a fixed value of 0x47 identifies the start of the packet. The header contains packet identification, scrambling, and control information. A continuity counter increments for each packet and allows a missing or out-of-sequence packet to be detected.

The header is followed by 184 bytes of MPEG-2 or auxiliary data. Optionally, adaptation headers may be inserted to carry specific information that is not required in every MPEG-2 packet, for example, the program clock reference (see upcoming subsection on Timing Distribution). The Packet Identifier (PID) has local significance within the system multiplex: It makes it possible to demultiplex the transport stream at the receiver and to select the program of interest. Alternatively, one program can be processed for display while another one is recorded for later use. The programs recorded do not need to be further processed. That can take place at playback.

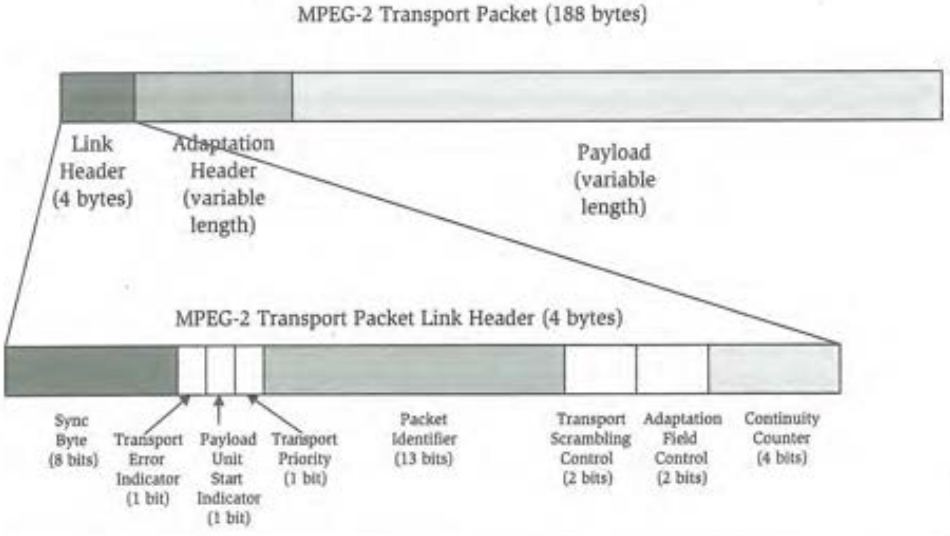


Figure 3.22 MPEG-2 Transport Packet Format. (From [39] used with permission)

The system multiplex can be extremely complex. Individual packetized elementary streams can be multiplexed together to form a program that includes multiple audio streams, multiple video streams, and ancillary signals such as conditional access control messages. This is called a *single-program transport stream* (SPTS). If the data capacity of the channel is sufficient to carry more than one SPTS, they can be multiplexed into a multiple-program transport stream (MPTS). In this way, many programs can be time division multiplexed into a transport stream of an arbitrary bit rate. The entire transport stream is then modulated onto a carrier signal to form a "digital" channel. Many channels can be combined to form a frequency division multiplex of several hundred megahertz.

Figure 3.23 shows an alternative view of a system multiplex. The bottom layer is the forward application transport (FAT) channel, which contains the MPTS. It is divided into PES packets, carrying compressed audio and video bitstreams, and program-specific information (PSI), carrying tables that describe the system multiplex. The next subsection, on Navigation, will describe how these are used by the receiver to find and select the desired program.

Navigation

Although the MPEG-2 transport stream syntax includes some basic structures to enable the receiver to select one out of many programs included in an MPTS, these structures are not sufficient to support the navigation features demanded by a sophisticated receiver. Moreover, they do not provide information to support an electronic program guide (EPG). We will describe the structures defined by the MPEG systems layer before moving on to necessary extensions introduced by the SCTE and the ATSC.

The MPEG-2 system multiplex always includes a program association table (PAT), which is transmitted continuously on PID value 0. The PAT contains an entry for each program within the multiplex and information identifying the PID

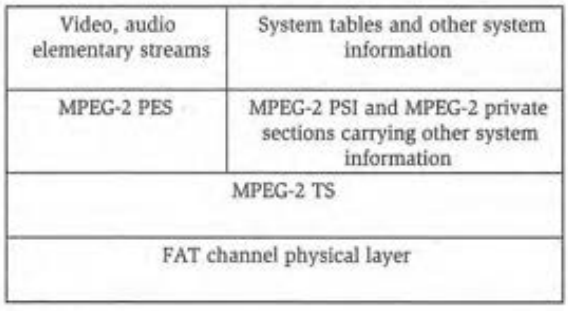


Figure 3.23 FAT Channel Transport Layer Protocol. (From SCTE 40 2001)

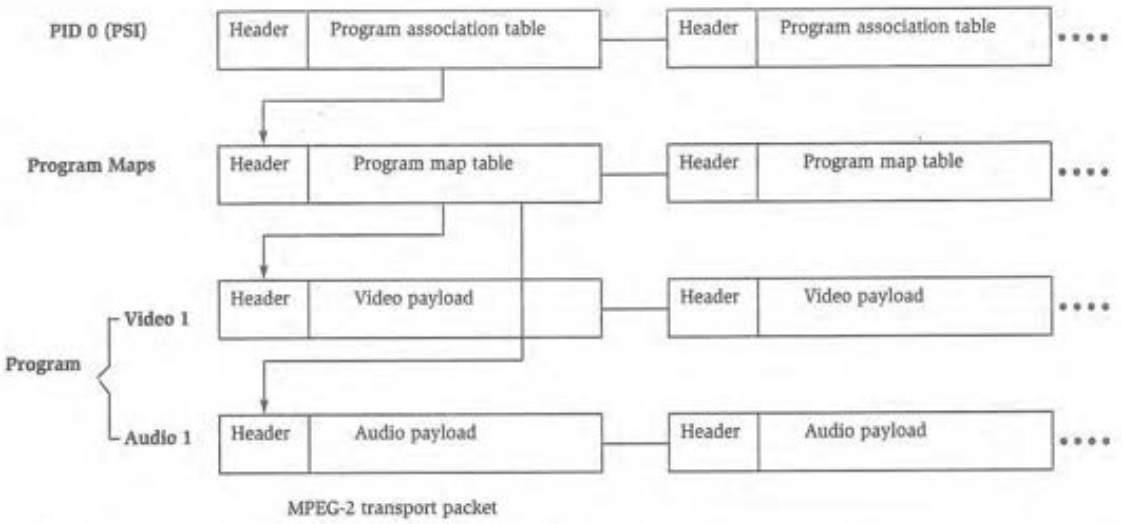


Figure 3.24 Program Association and Program Map Tables. (From [40] used with permission)

of the program map table (PMT) for it. In turn, the PMT contains the video PID(s), audio PID(s), and associated data PID(s) for the program.

Figure 3.24 illustrates how a receiver "navigates" the PAT and PMT to discover the video and audio PESs that make up the program. However, this is a very basic form of navigation; there is no description of the program name (for example, "PBS") or the program description (for example, "World News Hour").

In cable systems, information about the frequency and description of each channel is typically sent in a separate out-of-band data channel. This information is called the *channel map*. When digital transmission was introduced, it was a simple extension to add information to describe the new digital channels, such as specifying the program number as well as channel frequency. In addition, as electronic program guides were developed, additional information was sent to the STT to carry guide data. The format of the descriptions is unique to each vendor.

When terrestrial digital broadcast was introduced, the ATSC had to standardize a single descriptive format and include it in each system multiplex, because there is no equivalent of an out-of-band data channel for terrestrial broadcast. ATSC standard A/65 is called Program and System Information Protocol (PSIP).[41] The following tables are included in PSIP and shown in Figure 3.25.

System time table (STT) — provides the system time of day.

Master guide table (MGT) — defines other, optional tables, for example the EIT and ETT tables.

Virtual channel table (VCT) — provides information about the virtual channels (programs) contained in the system multiplex.

Rating Region Table (RRT) — provides a definition of the content advisory ratings for the system multiplex.*

All of these tables are transmitted on a well-known PID known as the *base PID*, which is defined as 0x1FFB for PSIP. Additional tables can be defined on separate PIDs, the event information table and extended text table are used to transmit further information about the programs in the system multiplex:

Event information table (EIT) — lists upcoming events that will be broadcast for each virtual channel (program) for the next 3 hours. Up to 128 EITs can be sent, each on its own PID, providing upcoming information for up to 16 days. The first four EITs are mandatory.

Extended text table (ETT) — lists descriptive information about each virtual channel (program). Just like the EITs, they have a 3-hour window, and up to 128 ETTs may be sent. However, they are optional.

Now that we have described the system information for broadcast, we can discuss the methods for cable systems. The ATSC assumed that cable systems, as rebroadcasters of off-air content, would naturally follow the A/65 standard. However, much of the programming on cable is no longer rebroadcast, and cable had already developed proprietary mechanisms for the carrying and coding of navigation information. Nevertheless, a compromise had to be worked out if it was ever going to be possible for a manufacturer to build a cable-ready digital television (CR-DTV). (See Section 3.7.4)

The approach was to take out-of-band information and make it available to the digital TV via the POD–host interface (see Chapter 21). The format is very

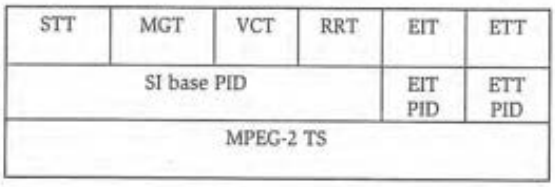


Figure 3.25 Protocol Stack for In-Band System Information. (From SCTE 40 2001, used with permission)

* The RRT for the United States is "hardwired" for two reasons: (1) It doesn't follow the linear "scale" of the original design and thus is a special case, and (2) since it is hardwired, the actual transmission of the RRT is not needed unless changed. Other regions can have downloaded RRTs and dynamic content advisory descriptors to match.

similar to the A/65 standard and allows an operational migration path to full A/65 compliance over time. (See ANSI/SCTE 65 Annex C for more details.)

Timing Distribution

The MPEG-2 transport stream carries timing information by means of a time stamp called the *program clock reference* (PCR). A PCR is placed in the adaptation header frequently enough (typically 10 times/second) to allow a phase-locked loop to recover the system time base at the receiver.

Video and audio PESs contain their own timing information, in this case presentation and decode time stamps. These are used by the decoder to synchronize the audio to the video.

Conditional Access

Although the MPEG-2 systems layer does not specify any conditional access (CA) mechanism, it does provide "hooks" for including conditional access control information in the system multiplex. Conditional access is not required and may be omitted if the multiplex is to be sent in the clear.

If the MPEG-2 system multiplex includes a conditional access table (CAT), it is transmitted continuously on PID value 1. The CAT contains an entry for each program within the multiplex, and the PID of the entitlement management message (EMM) stream for the CA system. (In cable systems, EMMs are typically sent out of band, and so this PID is usually null.)

In addition, the PMT may contain the PID for an entitlement control message (ECM) stream for the program. ECMs contain encrypted control words for the decryption circuitry (see Chapter 21).

Figure 3.26 illustrates how a receiver uses the CAT and PMT to obtain the information it needs to descramble the program.

Splicing

In a compressed video, bitstream cutting, or splicing, from one feed to another is not trivial. There are two main reasons for this.

1. The group of pictures (GOP) structure (see Figure 3.16) defines the frame order for a particular compressed bitstream. It defines which frames (or pictures) form a reference for predicted frames. If the bitstream is changed at an arbitrary point, these references will no longer make sense.
2. The encoder maintains a model of the decoder buffer fullness and codes each frame within a bit limit to ensure that the decoder buffer stays within prescribed bounds. If the bitstream is changed, the buffer will likely overflow or underflow.

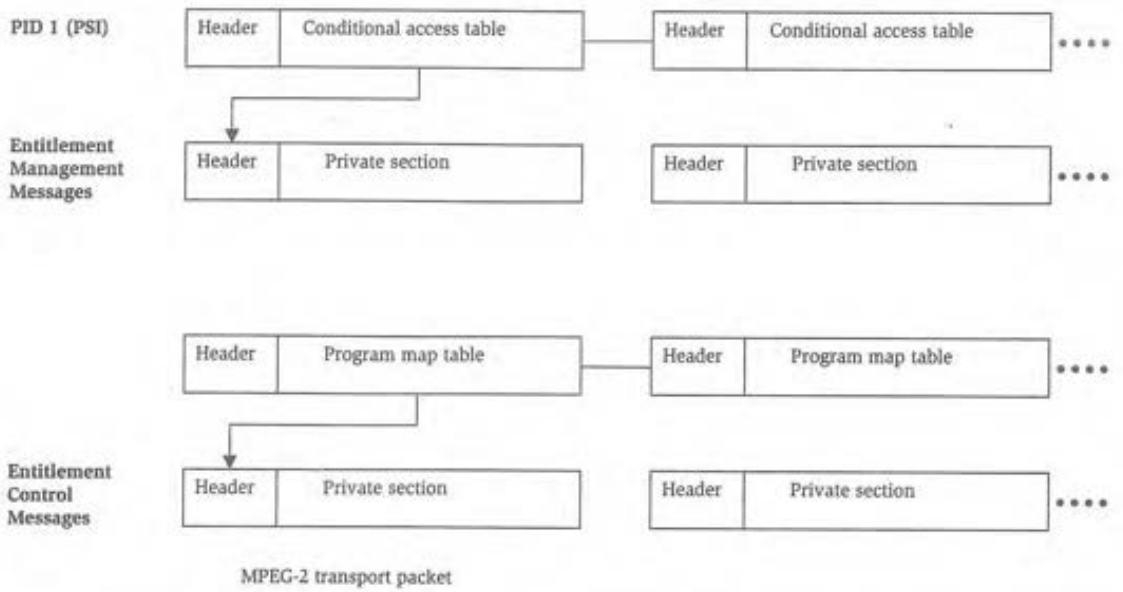


Figure 3.26 Conditional Access Table. (From [42] used with permission)

The Society of Motion Picture Experts (SMPTE) have defined mechanisms to establish a splice point in an MPEG-2 transport stream such that a seamless splice can occur.[43] At a splice point, the first frame of the new feed will be an I frame, and the GOP structure will specify that the predicted frames of the old feed should not use it as a reference frame. In addition, the buffer fullness is defined at the splice point so that the new feed can be encoded to fit within those constraints. For a predetermined splice, a fade to black at the splice point makes the transition much easier because there is no information to encode; still, it is possible to do a seamless splice of full-motion video. (However, note that this is not the method adopted by cable for digital program insertion — see Section 3.5.4).

Similar mechanisms exist for audio bitstreams, but splicing is much less complicated and audio muting is usually employed to cover any hiccups due to the transition.

3.5.2 Statistical Multiplexing

The goal of digital video compression is to reduce wasteful redundancy in the programming signal. Since programs have different amounts of redundancy at any given time and since the amount of redundancy varies over time in a given program, the compressed bitstream will have an instantaneously variable bit

rate. To accommodate this, the receiver needs to have sufficient buffer memory to smooth the flow of data to the decoders. This introduces a latency to accommodate buffer filling and emptying, and it requires standardization on the minimum buffer size. Although the buffer memory allows some flexibility, the problem of managing the flow of data is not completely eliminated. To ensure that the receiver buffer never underflows or overflows, the encoder maintains a model of the receiver buffer fullness and regulates the encoding rate so that the receiver buffer fullness always remains within design limits.

When the capacity of the channel compared with the average data rate of the programming allows multiple programs to be carried, an opportunity results. When there are multiple programs, the likelihood is that there will be a distribution of complexities among the channels that is amenable to statistical analysis. This is a reapplication of a phenomenon applied by long-distance phone companies decades ago. It was determined that much of conversation involves pauses when nothing is being said. If a large number of conversations are combined, the dead time of one conversation could be used to carry other conversations. If half a dozen conversations are combined, there is sufficient statistical dead time for the addition of one or two more conversations. Efficiency is dramatically increased. The scarce resource of transmission capacity is better utilized. The larger the number of combined conversations, the better the statistics and the better the results. This process is called *statistical multiplexing*. It applies equally well to video signals that have varying amounts of relatively dead time. The dead time in video is the time when little changes from frame to frame and the need for transmission capacity is minimal.

In practice, excellent results can be obtained with statistical multiplexing of video if there is a reasonably large number of channels within the system multiplex. For example, 10 constant-bit-rate 3.5-Mbps video bitstreams may be replaced by 14 video bitstreams with the same total instantaneous bit rate of 35 Mbps, producing a statistical gain of 40% with no perceived reduction in quality. However, it is important that the video bitstreams be uncorrelated; that is, they cannot all switch to challenging sports material at the same time.

Buffer memories for each of the multiplexed bitstreams facilitate the process of statistical multiplexing but require sophisticated algorithms to manage their utilization. As a buffer memory reaches its capacity, provisions must be made for it to be relieved. The better the algorithm, the more effective the result. The management of the buffer memories can be done at the origination point. As better algorithms become available, better results can be obtained using the existing receiving hardware.

Statistical multiplexing efficiency can be optimized if a bank of real-time video encoders is combined with a statistical multiplexer. In this arrangement, each encoder dynamically requests a bit rate according to the complexity of material at that point and a specified quality level (as set by an operator). The statistical multiplexer will dynamically assign a target bit rate to each encoder, depending on (a) the bit rate requested by it and (b) the sum of the bit rates

requested by all encoders. In this way, each video encoder will achieve the best quality it can, given the constraints established by the multiplexer. In addition, statistical multiplexing systems may incorporate "look-ahead" preprocessing to anticipate bit needs before they happen. Using these techniques, the multiplexer is able to use the entire capacity of the channel with no possibility of overflow.

3.5.3 Reencoding and Transcoding

Reencoding provides a way to dynamically change the encoding parameters of a video stream after it has been encoded. Various techniques have been developed, but fundamentally they all work by requantizing the DCT coefficients. In some cases a significant reduction in bit rate can be achieved with little or no loss of quality.

Transcoding can be used to convert a bitstream encoded in one format to another. This can be done by decoding to baseband video and then encoding according to the new format; however, this is expensive and causes delay. Worse still, the second encoder may struggle to encode artifacts created by the first because they contain significant random information (for example, blocking or mosquito noise). A well-designed transcoder can minimize these effects, and reduce delay, by recoding the bitstream without the need to decode all the way back to a baseband signal. Better results are possible because it knows about the source and destination formats. Transcoders are able to change resolution and even render a graphical overlay on the video.

Clearly, reencoding or transcoding is an important tool at any point in which compressed bitstreams are combined to create a system multiplex. Statistical multiplexing can be much more efficient because peaks that occur simultaneously across all bitstreams can be smoothed by judicious buffer management, requantization, or even transcoding. In addition, reencoders and transcoders maintain a model of the decoder buffer fullness and work within its constraints so that the receiver buffer will not underflow or overflow.

3.5.4 Digital Program Insertion

Local commercial insertion is an important source of revenue for the cable operator. It is simple to cut from one audio/video feed to another, and the practice of *local ad insertion* has been commonplace for many years, originally using videotape machines to play the local commercials and more recently using *digital ad insertion* servers. The time periods allotted for the insertion of commercials, known as *avails*, are signaled by cue tones, which are sent about 5–8 seconds before the commercial is to be inserted (allowing for *preroll* of the tape machine or server). (See also Section 8.9)

In a digital headend, equivalent functionality must be preserved; that is, it must be possible to insert a digitally compressed commercial into a MPEG-2 transport

stream. It is theoretically possible to decode all the programs in an MPTS to baseband video and audio, insert the commercial using traditional techniques, and then reencode and reassemble the MPTS. However, this approach is not recommended because it would be extremely expensive and cause a significant time delay and, very likely, a loss of quality. Therefore, an SCTE Digital Video Standards subcommittee has worked very hard for a number of years to develop standards for *digital program insertion* (DPI) based on splicing of compressed digital content[44,45].

The DPI standards do not specify exactly how the splicing of the compressed bitstream should be achieved (see Section 3.5.1), because there are a number of acceptable techniques, and the intention is to allow innovation on the part of the vendors. Encoding and/or transcoding methods are extremely useful to allow a smooth transition between one bitstream and another; in practice, many DPI splicers use these techniques to allow near-seamless splicing at an arbitrary point in the bitstream. Instead of specifying the splicer implementation, the DPI standards specify a standard *applications programming interface* (API) between the network and the DPI system. The reference model for DPI is shown in Figure 3.27.

The input to the DPI system, called the *network stream*, is an MPEG-2-compliant transport stream. Any programs that are to be spliced will include a *registration descriptor* in their PMT. In addition, a separate PID stream carries a *splice information table*, which contains the splice commands and the presentation time of the splice points for that program.

The splice subsystem intercepts and, if necessary, decrypts the splice commands before passing them to the local ad server. At an avail, the ad server plays out the MPEG-2 audio/video elementary streams and the splicer inserts them

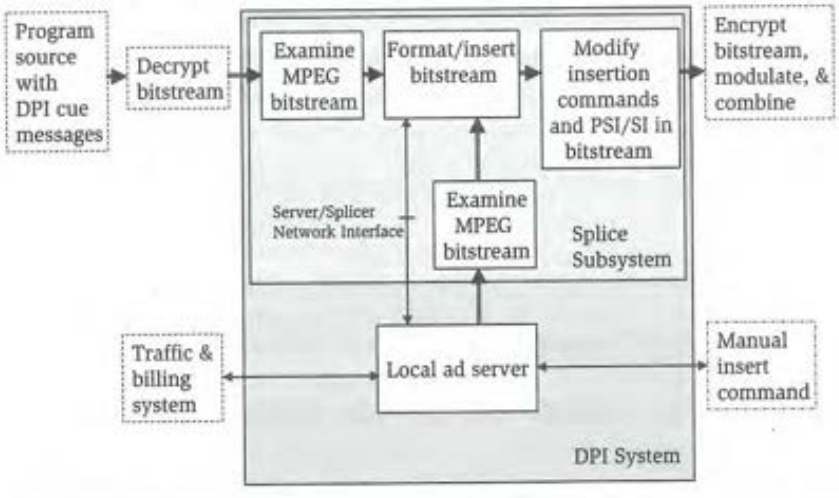


Figure 3.27 DPI Headend System Overview. (From [46] figure 1, page 5, used with permission)

into the network stream from the specified *splice in-point* until the specified *splice out-point*. An NCTA conference paper provides a good background on some of the implementation details of DPI for further reading.[47]

3.5.5 Emergency Alert System

Local emergency alerts are currently provided by a crawling text overlay along the bottom of the screen. Sometimes they are accompanied by audio tones. In other applications, text and graphics are overlaid on the video. During a national alert or certain significant local/regional events, a video and audio override of every channel is required. The common analog practice of source switching is relatively simple to implement, but it is much more expensive and complicated to do this with a digital multiplex.

The SCTE have defined an emergency alert message for cable to allow the alert event to be signaled to the receiver.[48] All digital receivers have the capability to render text messages over the video and to generate audio tones. In the unlikely event of a national emergency, the receiver can be instructed to force-tune to a channel for the duration of the alert. The emergency alert message is formatted as an MPEG-2 table that contains the following fields:

- Start time and expected duration of the emergency alert
- Text message description of the alert
- Availability and location of a details channel
- Force-tune flag
- A pointer to an audio channel to replace the audio for the duration of the emergency

The emergency alert message may be sent in-band within the system multiplex for in the clear channels, allowing it to be inserted by a terrestrial digital broadcaster for retransmission over a cable system. Alternatively, for scrambled channels, the emergency alert message can be transmitted in the out-of-band channel. In this case, it is received by the point of deployment (POD) module and relayed to the host as an MPEG-2 table (using the same mechanism previously described for navigation messages).

3.6 Digital Transmission

Now that we have described how an MPEG-2 system multiplex is created, we need to explore how it is delivered over an analog cable system. Moreover, the MPEG-2 transport stream assumes an error-free transmission path, and so forward error correction information must be added to the multiplex before it is modulated onto a carrier signal.

3.6.1 Modulation

Modulation serves the same purpose for digital signals as it does for analog signals. Although digital signals can conveniently utilize time division multiplexing, frequency division multiplexing (FDM) remains an important part of the plan. There are several reasons for this. Practically, digital circuits are limited in their speed. They cannot support a TDM structure for the entire gigahertz spectrum of cable. Breaking the spectrum into 6-MHz pieces relaxes the speed requirement on the digital circuits. TDM can then be utilized within each of the 6-MHz channels.

Additionally, FDM facilitates the transition from analog television to digital television. Cable systems have added digital for those customers who want more programming choices. They have added digital to the current collection of analog channels. In some cases, the analog scrambling is left in place, but it is also common to simulcast premium channels in the digital tier to avoid the cost of incorporating analog descrambling circuitry into the digital STT. By separating the spectrum into two areas, one for analog channels and one for digital channels (including TDM), a rational transition to digital technology has been accomplished.

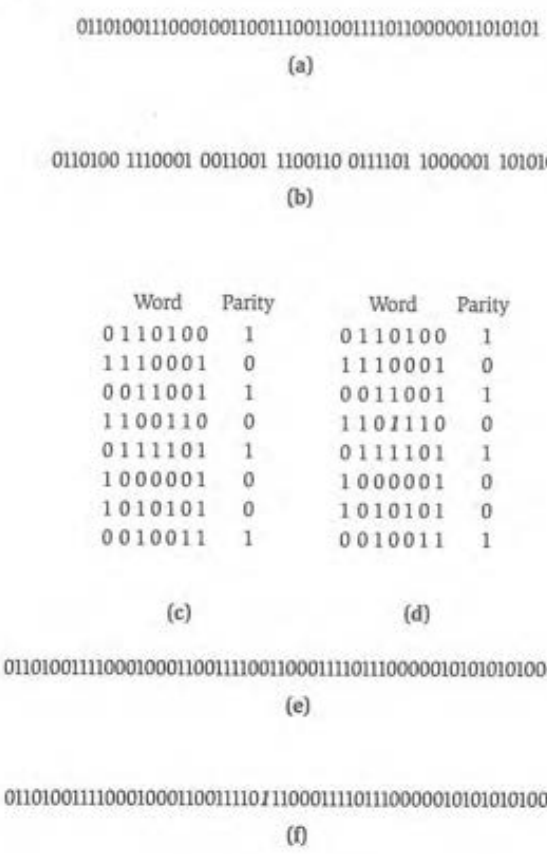
As the cost of a digital STT continues to fall, it is becoming feasible to use digital compression as an alternative to a plant upgrade. For example, in a 450-MHz system, only must-carry channels are sent in analog format (maybe channels 2–40), and all other programming is sent in digitally compressed format. Each subscriber would need a digital STT, but the cost may be less than that for upgrading the system to 750 MHz.

3.6.2 Forward Error Correction (FEC)

One of the most important features of digital signals is the ability to judiciously add redundancies that allow error detection and correction. The understanding of error detection and correction begins with the realization that an error is the conversion of a 1 to a 0 or the conversion of a 0 to a 1. The fact that the error is limited to just those two cases is very powerful. Earlier in this chapter, we discussed the simplest of error-detection schemes, parity checking. If parity is taken into two dimensions, substantially more powerful results can be obtained.

Figure 3.28(a) shows seven binary words of seven bits each. In Figure 3.28(b), spaces are introduced to help identify the clusters of seven bits. In Figure 3.28(c), the seven binary words of seven bits each are arranged in a two-dimensional array. An even parity bit is attached to each row. Recall that even parity means that the appended bit will be a 1 if an odd number of 1s appear in the word and a 0 if an even number of 1s appear. Parity allows us to determine if an odd number of errors has occurred in any of the words, but it does not allow us to determine where the error occurred.

Now let us form parity bits in the vertical direction as well. This will give us an eight-bit byte at the bottom of the array. We now have 8 columns and 8 rows, for 64 bits. This array can be strung out into a serial stream, as shown in Figure 3.28(e).



01101001110001001100111001100111011000001101010 1

(a)

0110100 1110001 0011001 1100110 0111101 1000001 1010101

(b)

Word	Parity
0110100	1
1110001	0
0011001	1
1100110	0
0111101	1
1000001	0
1010101	0
0010011	1

(c)

Word	Parity
0110100	1
1110001	0
0011001	1
1101110	0
0111101	1
1000001	0
1010101	0
0010011	1

(d)

01101001111000100011001111001100011110111000001010101010001001 11

(e)

01101001111000100011001111011100011110111000001010101010001001 11

(f)

Figure 3.28 Block code example.

If an error occurs, such as shown in Figure 3.28(f), it will convert a 0 to a 1 or it will convert a 1 to a 0. In this example, a 0 has been converted into a 1. If at the receive site, we rearrange the received bits into a block as shown in Figure 3.28(d) and apply the parity-checking principle to the rows and columns, we will find a violation in row 4 and column 4. This tells us we have an error and it also tells us where! Since an error can only be an inversion of the logic, we can correct the error perfectly. A little experimentation with this example will reveal that if we have two or more errors that do not occupy the same row or column, we can locate them and correct them. This would work for up to seven errors as long as they were not in the same row or column. If, however, two errors occurred in the same row (or the same column), we could determine that we have two errors, but we could not locate the row (column) in which they occurred. Knowing that two errors exist is valuable information.

In this example, a penalty is paid in complexity. Thirty percent more information is sent. In addition, the hardware at the sending and receiving end is more complex. This example is termed a *block code*, since the data is partitioned into blocks for processing. More complex codes exist that are beyond the scope of this text to explain. They involve advanced mathematical concepts and are very effective.

Several error-combating strategies are often combined in digital transmission systems. A most important error-combating strategy is called *interleaving*. Figure 3.29 shows how this works. The goal is to take adjacent bits and spread them out in time so that a noise burst does not destroy adjacent bits. When the

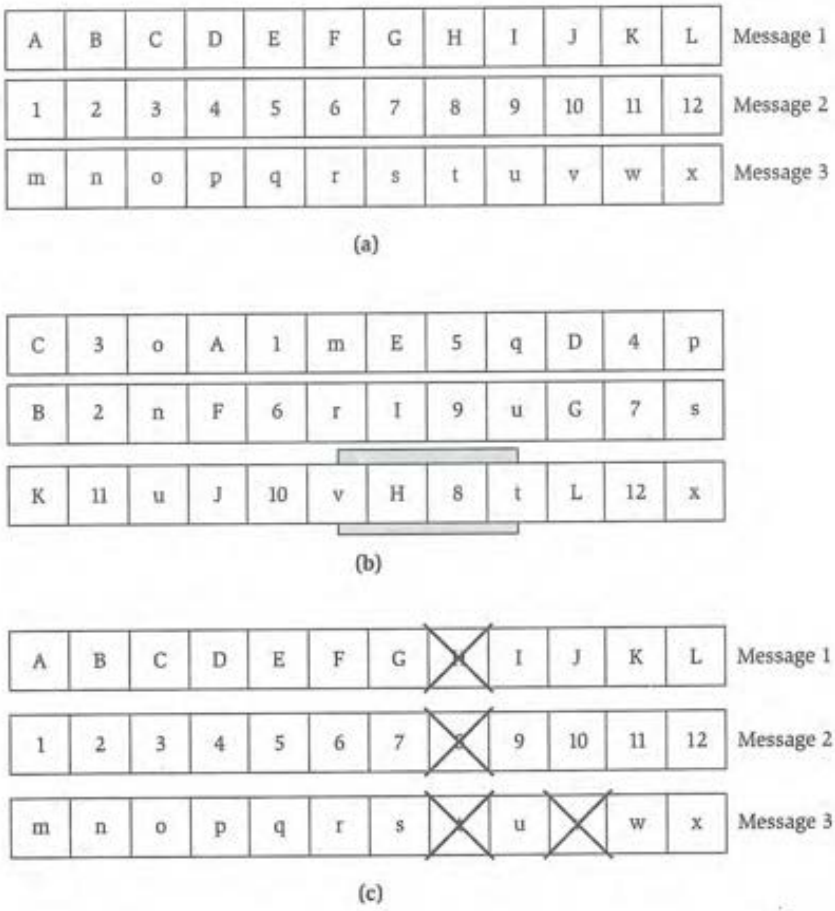


Figure 3.29 Interleaving to protect against burst errors. (a) Order of interleaving. (b) Interleaved messages suffering a burst of noise. (c) De-interleaved messages suffering nonadjacent losses.

interleaving is undone in the receiver, the group of destroyed bits is distributed over a length of time, so damaged bits are not adjacent. Then techniques such as the block code example can detect and correct the errors. Interleaving converts bursty noise into something more like random noise.

3.6.3 Required Signal Quality

Digital systems have a "cliff effect" that allows them to have perfect or near-perfect pictures as the received signal weakens, until a point is reached where the picture fails altogether. It happens very abruptly. All viewers have essentially the same quality picture and sound, no matter where they are in the service area. When the signal strength falls below a threshold, the picture essentially disappears.

Noise in an analog picture appears as snow and begins to mask the image and get in the way of the enjoyment of the picture. If the programming is compelling, viewers will put up with this and continue watching even very poor pictures. If the programming is marginal, viewers will switch to something else. Viewers of digital images won't get the chance to make that decision. The picture will be just as noise free at the home as in the studio. When the noise becomes excessive, the picture will be useless. It will fall off the cliff.

Multipath distortions (ghosts) that degrade an analog picture either destroy the digital picture by pushing it over the threshold or have no impact at all. The digital broadcast system's built-in multipath canceller system brings reception to areas not previously served.

Field tests in North Carolina have indicated that the broadcast digital system is so robust that anywhere an analog picture was even marginally acceptable, the digital picture was perfect. In many cases where the analog picture was unwatchable, the digital picture came in just fine.

Cable spectrum is much better behaved than the over-the-air broadcast media. Either the ghost canceller is not needed or its design parameters are much more relaxed as compared with the broadcast case. If the analog signal is minimally acceptable, the digital signal will provide essentially perfect pictures (to the extent that the compression quality is acceptable).

3.6.4 Picture Quality

Picture quality is a function of the residual digital artifacts. These are determined primarily by the aggressiveness of the compression, the detail and motion in the picture, and the quality of the encoder.

Although expert viewers can easily spot digital artifacts in currently available digital video, there is little to no complaint from consumers. Standard-definition movies encoded at 3 Mbps using MPEG-2 MP@ML provide excellent results (by comparison, the average bit rate of a DVD is 4 Mbps). Movies can be encoded in high definition at 12 Mbps at MPEG-2 MP@HL with almost no loss of quality.

Video can be more challenging, especially high-motion material such as sports events. Variable-bit-rate compression combined with statistical multiplexing is commonly used to keep up with rapid motion while maintaining an efficient multiplex. To compress difficult sports material up to 6 Mbps may be required during peaks, even with a good MPEG-2 encoder. However, an average rate of 3 Mbps will probably suffice for a system multiplex containing a range of video material (sports, news, current affairs, and so on).

3.7 Digital Television Standards

Standards are of vital importance to the deployment of digital television. A digital STT is extremely complex and can be manufactured cost-effectively only in very high volumes. Moreover, standards support the development of high-volume chipsets that can be sold to a worldwide marketplace.

In a terrestrial broadcast environment, a standard format for digital television is required that is understood by all receivers that can receive the signal. Hence, the ATSC standard had to be developed in advance of deployment. In cable (and satellite) environments, proprietary systems were developed initially, but even in this case common standards (MPEG-2 video compression and transport) have been used to build cost-effective systems.

3.7.1 ATSC

The Advanced Television Systems Committee (ATSC) has developed a Digital Television Standard[49] that is a complete system for terrestrial broadcast of high-quality video, audio, and ancillary data. It defines the following.

1. Source coding and compression: MPEG-2 video compression (standard and high definition) and AC-3 audio compression (A/52).
2. Service multiplex and transport: The MPEG-2 systems layer is used for multiplexing and transport.
3. RF/transmission is 8-VSB (terrestrial mode) or 16-VSB (high-data-rate mode).
4. Navigation of program content within the service multiplex is defined by the Program and System Information Protocol (A/65).

For further information, the reader is referred to the *Guide to the Use of the Digital Television Standard*.[50]

3.7.2 SCTE

The Society of Cable Telecommunications Engineers (SCTE) is an accredited American National Standards Institute (ANSI) organization that utilizes the due

process methods required by ANSI to set standards for the cable industry. This is the mechanism by which digital video for use on cable is standardized. These SCTE standards are an integral part of OpenCable.

The *Digital Video Systems Characteristics Standard* covers the subset of MPEG-2 video compression adopted by cable.[51] The *Digital Video Transmission Standard* for cable television describes the framing structure, channel coding, and channel modulation for a cable television realization of a digital multiservice distribution system.[52] The standard applies to signals that can be distributed via satellite directly to a cable system, as is the usual cable practice. The specification covers both 64-QAM and 256-QAM.

The modulation, interleaving, and coding are tailored to the characteristics of cable system practice in North America. The modulation is selected by the cable system to be either 64-QAM or 256-QAM. The forward error-correction approach utilizes a concatenated coding approach that produces low error rates at moderate complexity and overhead. The system goal is one error event per 15 minutes. The input data format is MPEG-2 transport packets, which are transmitted serially with the MSB first.

Two modes are supported: Mode 1 for use with 64-QAM with a symbol rate of 5.057 Msps, and Mode 2 for use with 256-QAM with a symbol rate of 5.361 Msps.

The cable channel, including optical fiber, is a bandwidth-limited linear channel, with a balanced combination of white noise, interference, and multipath distortion. The QAM technique used, together with adaptive equalization and concatenated coding, is well suited to this application and channel.

3.7.3 OpenCable

CableLabs, the cable industry's research and development consortium, has created a number of purchasing specifications on behalf of the industry. These are not standards, since CableLabs is not an accredited standards-setting organization. Rather, these specifications define the functionality and external parameters of products used in cable but leave the internal details to the manufacturers. For example, CableLabs will not specify a single microprocessor or operating system for these applications. Nevertheless, to ensure interoperability, CableLabs establishes a certification program to verify that the products comply with the specification and will perform satisfactorily.

CableLabs announced a "Harmony Agreement" for digital standards in October 1997. Five open-layer standards are included in the agreement: MPEG-2 video compression, Dolby Digital (AC-3) audio system, MPEG-2 transport, ATSC system information, and the International Telecommunications Union ITU-J.83 Annex B modulation system. The modulation system calls for 64-QAM carrying 27 Mb/s and 256-QAM carrying 38.8 Mb/s with concatenated Trellis code plus enhancements such as variable interleaving depth for low latency in delay-sensitive applications. All these agreements have been endorsed by CableLabs

and have since been adopted as digital standards by the Society of Cable Telecommunications Engineers.

The goal for OpenCable is to create a multivendor interoperable environment of digital set-top terminals. An important objective is the competitive supply of all the elements of the system, including the advanced set-top terminal, its operating system, core applications and plug-ins, conditional access system, platform management system, and billing, provisioning, and subscriber management systems.

The procedure began with a request for information (RFI) sent to leading computer and consumer electronics companies. Twenty-three companies responded. From these submissions, CableLabs solicited vendors to author specific elements of the suite of specifications. An intellectual property (IP) pool was created with cross-licensing of all IP required for the specification. Implementation IP not inherently required by the specification was left to vendors to license in their normal fashion.

Key OpenCable interfaces include the formats for digital cable television, consumer privacy systems, copyright protection system, interfaces for high-speed connection to the Internet, and interfaces required to author interactive applications. The last of these is the OpenCable Applications Platform (OCAP), which is designed to allow the cable operator to download applications to retail STTs in order to provide a seamless environment for the cable subscriber. OCAP is a middleware specification that enables application portability across a wide range of home devices and cable networks. OCAP 1.0 defines a Java-based execution engine (EE), and OCAP 2.0 extends that platform with the addition of Web-based technologies such as XHTML, XML, and ECMAScript.

CableLabs will implement a vendor's certification process, whereby specific products can be tested for compliance with the OpenCable specifications. The certification process will allow these devices to be made available to consumers through retail outlets. CableLabs recommends that its members utilize the specifications for their procurements. However, members will make their own procurement decisions. This is not a joint purchasing agreement. The OpenCable specifications are available through the OpenCable Web site (www.opencable.com).

3.7.4 Digital Cable Compatibility

There has been a lot of interest in making "digital cable-compatible" devices available at retail. Cable operators and consumer electronics manufacturers both stand to gain, and the FCC believes that competition among manufacturers will provide more and better choices to the consumer.

A memorandum of understanding (MOU) between eight larger MSOs and 14 consumer electronics manufacturers was sent to the FCC in a letter to the chairman on Dec. 19, 2002. Although the memorandum covers only "unidirectional digital cable products," compliance with a significant number of standards is required to successfully tune, demodulate, descramble, decode, and display a

digital program. As already discussed, digital services tend to be "all or nothing"; that is, they work perfectly or not at all!

Although the MOU is just a first step toward cable-ready digital televisions, it is worth summarizing the list of SCTE standards that are required to support plug-and-play compatibility.

- The Cable Network Interface specification[53] establishes the interface between the cable plant and the cable-compatible receiver. Building on the Harmony Agreement, it includes RF transmission, modulation, video and audio compression, navigation, conditional access, Emergency Alert System (EAS) messages, and the carrying of closed captioning. Many of these specifications have already been discussed in this chapter and will be examined further in Chapter 21.
- The Digital Video Service Multiplex and Transport System for Digital Cable specification[54] defines all of the various operating parameters related to the MPEG-2 transport stream, including various constraints designed to ensure interoperability.
- The Service Information Delivered Out-of-Band for Digital Cable Television specification[55] describes how navigation information is sent over the cable system so that a receiver can display a list of available programs to the user and allow program selection.
- The Host–POD Interface specification[56] defines a standard interface between the cable-compatible receiver and a replaceable conditional access module (called a point-of-deployment, or POD, module). The cable operator will supply POD modules to subscribers to allow them to view premium digital channels. (See Chapter 22 for more details.)
- The POD Copy Protection System specification[57] defines a method of copy protection so that premium digital content passed across the POD–Host interface is protected from piracy. (See Chapter 22 for more details.)

The principle behind certification of cable-ready devices is primarily that they should do no harm to the network. A prototype receiver will be examined by CableLabs, but thereafter the consumer electronics manufacturers will self-certify their products. The provisions of the MOU are scheduled to take effect in 2004 for 750-MHz (or greater) cable systems, and further work is still required to enable advanced services such as enhanced program guides, video on demand, and interactive television applications.

3.8 Summary

We have discussed the fundamentals of digital compression technology for video and audio as they apply to cable television. We presented the discrete cosine

transform and its inverse transform. The DCT is applied to still images, and differential techniques are utilized to reduce redundancy between frames. For audio signals, sub-band and/or frequency transform coding may be used, combined with psychoacoustic masking, to substantially reduce bit rate while maintaining good perceptual quality.

Audio and video bitstreams are multiplexed into a single transport stream for distribution over the cable system. The transport stream includes information that describes the system multiplex, allowing the receiver to select and decode a single program. Error-detection and -correction techniques are used to maintain the integrity of the transport stream in the presence of a noisy transmission environment.

The application of digital techniques to cable dramatically increases the number of programs and the quality of the delivered signals while facilitating new services.

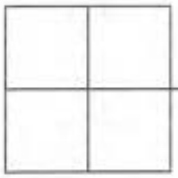
Endnotes

1. Chen and Pratt, Scene Adaptive Coder, *IEEE Transactions on Communications*, March 1984.
2. Joel Brinkley, *Defining Vision: The Battle for the Future of Television*. New York: Harcourt Brace, 1997, pp. 122–128.
3. Ibid., p. 116.
4. MPEG, *Information Technology — Generic Audio Coding of Moving Pictures and Associated Audio, Part 2: Video*. International Standard IS 13818-2, ISO/IEC JTC1/SC29 WG11, 1994.
5. Jerry D. Gibson et al., *Digital Compression for Multimedia — Principle and Standards*. San Francisco: Morgan Kaufmann, 1998.
6. Justin Junkus, *DigiPoints*, Vol. 2.: Society of Cable Telecommunications Engineers, 2000, Chap. 8.
7. Jerry Whitaker, *DTV: The Revolution in Electronic Imaging*. New York: McGraw-Hill, 1998, Chap. 5.
8. H. Nyquist, Certain Factors Affecting Telegraph Speed. *Bell System Technical Journal*, 3(1924):324–346.
9. H. Nyquist, Certain Topics in Telegraph Transmission. *AIEE Transactions*, 47(1928):617–644.
10. Jerry Whitaker, *DTV: The Revolution in Electronic Imaging*. New York: McGraw-Hill, 1998, p. 123, 189.
11. ITU-R BT.601-5, *Encoding parameters of digital television for studios*.
12. Mischa Schwartz, *Information Transmission, Modulation, and Noise*, 2nd ed. New York: McGraw-Hill, 1970.

13. S. Merrill Weiss, *Issues in Advanced Television Technology*. Boston: Focal Press, 1996.
14. IEEE, *Standard Specification for the Implementation of* 8 × 8 *Inverse Discrete Cosine Transform*, std. 1180–1990, Dec. 6, 1990.
15. Charles Poynton, *A Technical Introduction to Digital Video*. New York: Wiley, 1996.
16. K. Sayood, *Introduction to Data Compression*. San Francisco: Morgan Kaufmann, 1996.
17. D. A. Huffman, A Method for the Construction of Minimum Redundancy Codes, *Proceedings of the IRE*, Vol. 40, Sept. 1952, pp. 1098–1101. Institute of Radio Engineers, New York.
18. A. N. Netravali and B. G. Haskell, *Digital Pictures, Representation and Compression*. New York: Plenum, 1988.
19. EIA-608-B (2000), *Line 21 Data Services*.
20. ANSI/SCTE 21 2001 (formerly DVS 053), *Standard for Carriage of NTSC VBI Data in Cable Digital Transport Streams*, available at www.scte.org.
21. ANSI/SCTE 20 2001 (formerly DVS 157), *Standard Methods for Carriage of Closed Captions and Non-Real-Time Sampled Video*, available at www.scte.org.
22. EIA-708-B (1999), *Digital Television (DTV) Closed Captioning*.
23. SMPTE 240M (1995), *Standard for Television — Signal Parameters — 1125-Line High-Definition Production Systems*.
24. SMPTE 296M (1997), *Standard for Television*, 1280 × 720 *Scanning, Analog and Digital Representation, and Analog Interface*.
25. SMPTE 274M (1995), *Standard for Television*, 1920 × 1080 *Scanning and Analog and Parallel Digital Interfaces for Multiple-Picture Rates*.
26. ANSI/SCTE 43 2001 (formerly DVS 258), *Digital Video Systems Characteristics Standard for Cable Television*, available at www.scte.org.
27. MPEG, *Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to 1.5 Mbit/s, part 2: Video*. International Standard IS 11172-2, ISO/IEC JTC1/SC29 WG11, 1992.
28. Michael Adams, *OpenCable Architecture*. Indianapolis, IN: Cisco Press, 2000, Chap. 9.
29. J. Watkinson, *Compression in Video & Audio*. Boston: Focal Press, 1995.
30. K. Brandenburg and M. Bosi, *Overview of MPEG-Audio: Current and Future Standards for Low-Bit-Rate Audio Coding*. 99th Audio Engineering Society Convention, New York, Preprint 4130, 1996.
31. Jerry D. Gibson et al., op. cit., Chapter 8.
32. Jerry Whitaker, op. cit.

33. ATSC, *Digital Audio Compression Standard (AC3)*. Advanced Television Systems Committee, Washington, DC, Doc. A/52, Dec. 20, 1995.
34. MPEG, *Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to 1.5 Mbit/s, Part 3: Audio*. International Standard IS 11172-3, ISO/IEC JTC1/SC29 WG11, 1992.
35. MPEG, *Information Technology — Generic Audio Coding of Moving Pictures and Associated Audio, Part 3: Audio*. International Standard IS 13818-3, ISO/IEC JTC1/SC29 WG11, 1994.
36. MPEG, *MPEG-2 Advanced Audio Coding, AAC*. International Standard IS 13818-7, ISO/IEC JTC1/SC29 WG11, 1997.
37. MPEG, *Information Technology — Generic Audio Coding of Moving Pictures and Associated Audio, Part 1: Systems*. International Standard IS 13818-1, ISO/IEC JTC1/SC29 WG11, 1994.
38. Jim Taylor, *DVD Demystified*. New York: McGraw-Hill, 1998, p. 135.
39. Michael Adams, *OpenCable Architecture*. Indianapolis: Cisco Press, 2000, p. 64.
40. Michael Adams, *OpenCable Architecture*. Indianapolis: Cisco Press, 2000, p. 65.
41. ATSC, *Program and System Information Protocol for Terrestrial Broadcast and Cable*. Advanced Television Systems Committee, Washington, DC, Doc. A/65, Dec. 23, 1997.
42. Michael Adams, *OpenCable Architecture*. Indianapolis: Cisco Press, 2000, p. 66.
43. SMPTE 312M (1999), *Standard for Television, Splice Points for MPEG-2 Transport Streams*.
44. ANSI/SCTE 30 2001 (formerly DVS 380), *Digital Program Insertion Splicing API*, available at www.scte.org.
45. ANSI/SCTE 35 2001 (formerly DVS 253), *Digital Program Insertion Cueing Message for Cable*, available at www.scte.org.
46. ANSI/SCTE 67 2002 (formerly DVS 379), *Applications Guidelines for SCTE 35 2001*, available at www.scte.org.
47. M. Kar, S. Narasimhan, and R. Prodan, Local Commercial Insertion in the Digital Headend, *Proceedings of NCTA 2000 Conference*, New Orleans.
48. SCTE 18 2002 (formerly DVS 208), *Emergency Alert Message for Cable*, approved as a joint standard with CEA as ANSI-J-STD-042-2002, available at www.scte.org.
49. ATSC, *Digital Television Standard*, Advanced Television Systems Committee, Washington, DC, Doc. A/53, Sept. 16, 1995.
50. ATSC, *Guide to the Use of the Digital Television Standard*, Advanced Television Systems Committee, Washington, DC, Doc. A/54, Oct. 4, 1995.

51. ANSI/SCTE 43 2001 (formerly DVS 258), *Digital Video Systems Characteristics Standard for Cable Television*, available at www.scte.org.
52. ANSI/SCTE 07 2000 (formerly DVS 031), *Digital Video Transmission Standard for Television*, available at www.scte.org.
53. SCTE 40 2001 (formerly DVS 313), *Cable Network Interface Specification*, available at www.scte.org.
54. ANSI/SCTE 54 2002 (formerly DVS 241), *Digital Video Service Multiplex and Transport System for Cable Television*, available at www.scte.org.
55. ANSI/SCTE 65 2002 (formerly DVS 234), *Service Information Delivered Out-of-Band for Digital Cable Television*, available at www.scte.org.
56. ANSI/SCTE 28 2001 (formerly DVS 295), *Host–POD Interface*, available at www.scte.org.
57. ANSI/SCTE 41 2001 (formerly DVS 301), *POD Copy Protection System*, available at www.scte.org.



Chapter 4
Digital Modulation

4.1 Introduction

This chapter does not claim to be a treatise on data communications. Rather, it offers concepts that are particularly germane to the practice of data transmission on HFC networks. It is intended primarily to be an introduction to data communications for cable practitioners who have not previously been exposed to the subject. The examples of modulation methods and communications protocols are restricted to those that are most relevant to cable data communications.

Section 4.2 discusses elementary concepts in digital modulation. It is intended for engineers who may be familiar with analog modulation as practiced in the cable television industry but who have not previously worked with digital modulation. The section discusses the digital modulation formats most germane to cable television.

Section 4.3 introduces the concepts of multiplexing, a way of sharing the spectrum among many users. Section 4.4 introduces methods of measuring the quality of digitally modulated signals.

4.2 Modulation Technology

The object of data transmission is to get a stream of bits from here to there. *Bit*, which stands for *bi*nary digi*t*, is a voltage that can take on only one of two levels. We arbitrarily call these levels 1 and 0. Often, but not always, they are represented in electronic circuits by the voltages +3.3 (or +5) and 0 volts. The higher voltage usually, but not necessarily, represents a logical 1, while 0 volts represents a logical 0. A combination of many of these bits make up *digital data*.

It is possible to transmit bits of data over an RF carrier using the same variables as are used in analog transmission. One can vary, or "modulate," the amplitude, frequency, or phase of the carrier with the information to be transmitted. Frequency modulation is often used for very simple data communications links, due to the ability to construct low-cost FM receivers capable of very good performance. However, frequency modulation does not provide good "spectral efficiency."

4.2.1 Spectral Efficiency

One figure of merit used to compare different modulation formats is *spectral efficiency*, expressed in bits per hertz: the number of bits per second of data, divided by the RF bandwidth required to transmit it. A modulation type with high spectral efficiency can transmit more bits of data per second in the same number of hertz of RF bandwidth. On the other hand, it can be expected to cost more and be less *robust* (more susceptible to noise and transmission distortion). For example, we shall show that the theoretical spectral efficiency of biphase shift keying (BPSK) modulation is 1 bit per hertz. That is, when one is using BPSK modulation, the theoretical bandwidth required to transmit the signal is equal to the bit rate: a 1-Mb/s (one megabit per second) data signal will require 1 MHz of RF bandwidth. On the other hand, if the same signal were transmitted using 256-QAM, the theoretical bandwidth required is only 0.125 MHz because the spectral efficiency is 8 bits per hertz ($1/8 = 0.125$). However, all else being equal, 256-QAM requires a carrier-to-noise ratio (C/N) more than 13 dB better, and the cost of the hardware is considerably higher. In all cases, the actual occupied bandwidth is wider than the theoretical because of practical considerations.

4.2.2 Frequency Shift Keying (FSK)

The simplest commonly used modulation method is frequency modulation, called frequency shift keying, or FSK. Information is transmitted by shifting an oscillator between two frequencies. The receiver consists of a conventional frequency demodulator. It is followed by a shaping circuit, which restores the pulses to their original shape.

FSK systems tend to be popular for relatively low-speed data transmission where receiver cost is paramount. Applications include early-generation downstream data transmission to addressable set-top converters, upstream transmission from some early RF impulse pay-per-view (IPPV) converters, and status monitoring. Spectral efficiency is very poor, but this is relatively unimportant in low-speed applications. The transmission is robust (not easily damaged by noise or distortion), and the required hardware is quite low in cost. Single-chip data receivers have been available for many years.

Figure 4.1 illustrates a simple FSK communications system. Data, in the form of a series of 1s and 0s, is applied to a low-pass filter whose function is to minimize the number of sidebands produced by the modulation process. The filtered waveform is applied to a frequency modulated oscillator, whose frequency deviation is proportional to the voltage applied at the modulating input. *Frequency deviation* is the difference between an instantaneous frequency and the rest, or carrier, frequency. A parameter of an FSK system is the peak-to-peak deviation, or the difference between the two frequencies

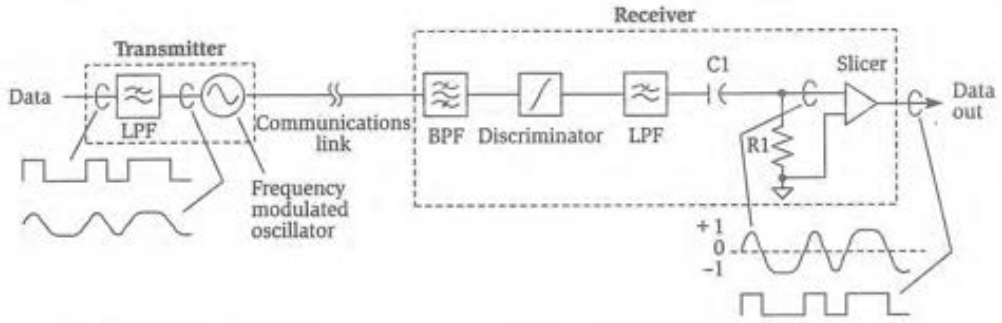


Figure 4.1 A simple FSK system.

of the oscillator when 1 and 0 are applied.* The wider the deviation, the more immune the system is to noise but the more spectrum the signal occupies.

After transmission to the receiving end, the signal is usually passed through a bandpass filter, BPF, and then applied to a discriminator, which is very similar to one used for analog signal recovery. A low-pass filter, LPF, is usually used after the discriminator, to reduce out-of-band noise picked up during transmission. Capacitive coupling is shown after the filter, to make sure the signal can be "sliced" at its average value. From the capacitor, the signal is supplied to a data slicer, a comparator whose output is 1 when the input voltage is above ground and 0 when it is below ground (or vice versa). The slicing process restores the wave shape of the original signal.

Most transmission systems require that the dc value of the data waveform be removed for transmission. This means that, in some manner, the waveform averaged over some reasonable interval must have a value of 0 volts. That condition is violated if the datastream contains an excessively long string of 1s or 0s. If the signal doesn't have the same quantity of 1 and 0 states when averaged over some suitable time interval, C1 will take on a charge and the slicing will not occur in the center of the waveform. Later we show several ways to ensure that a signal doesn't have a dc component.

* In the analog world, deviation is normally measured on a peak basis—the difference between either extreme frequency and the center frequency. Peak-to-peak deviation is twice the peak deviation. In digital transmission, the carrier never rests at the center frequency, so peak deviation tends to have less meaning. One must be precise, however, in distinguishing between peak and peak-to-peak deviation.

Filtering in FSK Transmission Systems

At least three stages of filtering are used in the transmission system. A low-pass filter at the transmitter limits the spectrum of the baseband signal (before modulation). In many other modulation methods, the filter may be implemented either at baseband or at RF, after modulation. However, in FSK the nonlinear nature of frequency modulation precludes doing so. The receiver contains a bandpass filter at RF (or intermediate frequency, IF), used to limit the energy applied to the discriminator. After demodulation, a baseband low-pass filter is used to remove as much noise as possible.

In low-cost FSK systems, the filtering philosophy often recognizes that either the transmitter or the receiver is in the subscriber's home, and this end must be built at minimum cost. Placing most of the filtering at the headend helps keep the cost of the home equipment down. This somewhat compromises the noise performance or bandwidth efficiency in favor of low cost. For example, a data communications system that communicates from the headend to addressable set-top converters will employ significant filtering at the transmitter. At the receiver, the IF filter frequently employs a low-cost ceramic filter made for FM receiver use (180- to 200-kHz bandwidth), and the deviation is set to make the occupied spectrum match the width of the filter.

Manchester Encoding

We now describe one method for adding a clock to the data signal and for removing any dc component. The method is independent of the modulation but is introduced at this point because it is often used with FSK modulation. Manchester encoding is also used in 10/100BaseT Ethernet. In data transmission, it is necessary not only to recover the transmitted data but also to recover a clock signal, used to define when a data state should be sampled. A clock is necessary in order to allow the receiver to "know" when a new bit arrives. The clock must be recovered from the transmitted data, which can be done in many ways. Manchester encoding is often used in low-cost systems due to its simplicity. The data is said to be "self-clocking." The penalty is that the spectral efficiency of the transmission is halved compared with what it would be otherwise.

Figure 4.2 illustrates one form of Manchester encoding, which is applied to the data before it enters the modulator. At the top of the figure is a string of data represented in NRZ (nonreturn to zero) format, in which the data bits appear one after another, with nothing in between. A separate clock signal is typically used to transfer NRZ data from one place to another. However, in a communications link, a separate clock path does not exist. Therefore, in some manner the clock must be embedded with the data. Manchester encoding provides a simple, albeit bandwidth inefficient, way to do this.

The center portion of Figure 4.2 shows the bit cell boundaries: the time when the signal transitions from one bit to the next. At the bottom of the figure

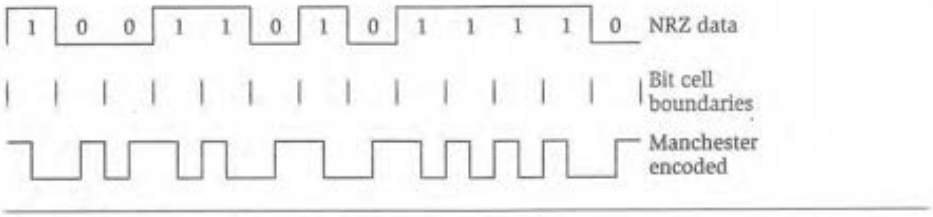


Figure 4.2 Manchester Encoded Data.

is shown the Manchester encoded data. In the middle of each bit cell, Manchester encoding causes the signal to transition from one level to the other. This regularly spaced set of transitions forms a clock for the receiver. The first half of each cell contains the value of the bit; the second half contains the complement (opposite state) of the bit (or vice versa). A variation is differential Manchester encoding: At the beginning of each cell containing a 0, the encoded signal makes a transition; at the beginning of each cell containing a 1, the transition is omitted. There is always a transition in the middle of the bit cell. On the other hand, if the transition occurs in the middle of a cell containing a 1 but not in a cell containing a 0, the encoding is called *biphase mark*. It is used in the S/P DIF interface described in Chapter 22.

Note that Manchester encoding is *not* a modulation method. Rather, it is a baseband coding method that is useful in low-cost systems to provide clock recovery and removal of a dc component of the data.

4.2.3 Baseband Spectrum of a Data Signal

The spectrum of a data signal may be understood with reference to the waveform of Figure 4.2 and the low-pass filtering of Figure 4.1. Figure 4.1 shows that it is possible (and usually desirable) to low-pass filter the data waveform so that when a string of alternating 1s and 0s (or a string of like bits in Manchester encoded data) is encountered, the waveform looks substantially like a sine wave. No useful data is lost in the filtering process, and the occupied bandwidth of the signal is thus minimized. Practical filters are somewhat wider than this, but the principle remains.

When a string of 1s and 0s appears in an NRZ signal (defined in Section 4.2.2.) and is optimally filtered, the resulting waveform is substantially a sine wave at a frequency of one-half the bit rate. To understand this, recall that a complete cycle of a sine wave occupies the time from one positive-going zero crossing to the next positive-going zero crossing. In order to produce such a waveform, two bits, 1 followed by 0, are required.

In general, a data signal can contain alternating 1s and 0s. This pattern defines the highest baseband frequency, equal to one-half the bit rate. However,

that data can also contain long strings of 1s or 0s, which represent a frequency approaching 0. Because of this, the fundamental spectrum of the baseband (unmodulated) data signal extends from one-half the bit rate down to very close to 0.

In summary, the spectrum of a digital signal extends from very low to at least one-half the bit rate (really the symbol rate, a concept to be introduced later in the chapter). Practical systems have frequencies that extend above this due to real-world filtering constraints. The excess bandwidth may range from a few percent to much higher, depending on the application.

4.2.4 Introduction to Bit Error Rate

An important figure of merit for any data communications system is the *bit error rate* (BER) of the system. This represents the proportion of the bits transmitted that are received incorrectly. If, on the average, two bits are in error for every million (10^6) bits transmitted, then the bit error rate is

$$\text{BER} = \frac{\text{number of bits received in error}}{\text{total number of bits transmitted}} = \frac{2}{10^6} = 2 \times 10^{-6} \qquad (4.1)$$

The BER is a function of the type of modulation, the characteristics of the modulator and demodulator, and the quality of the transmission channel. A common way to compare different modulation systems is to plot the BER against the signal-to-noise ratio (data communications practitioners tend not to differentiate between "signal" and "carrier" to noise). This is a convenient way to compare the capabilities of different modulation techniques and to predict performance in the presence of noise. This topic is covered in Section 4.2.12. Sometimes BER is plotted against signal-to-noise ratio in the presence of certain other impairments, such as group delay distortion, interference, and phase noise.

Obtaining an equation relating carrier-to-noise ratio to BER for an FSK system is difficult due to the nonlinear nature of the frequency modulation process. The literature contains only a few attempts, and contradictions exist. Within the parameters commonly used in cable television work (very wide deviation compared with the bit rate), the BER can be quite good at very low C/N. The penalty is that a significant amount of bandwidth is consumed, compared with that required for other modulation methods operating at the same bit rate.

4.2.5 Biphase Shift Keying (BPSK)

Biphase shift keying (BPSK) is the simplest of a class of modulation that is used extensively in cable television work. It is somewhat more complex to detect than is FSK, although producing a BPSK signal is simple. In a BPSK system, the phase is shifted 180° between transmission of a 0 and a 1.

Figure 4.3 illustrates a simple BPSK transmission system. An NRZ datastream is low-pass filtered (LPF1) to remove higher-order harmonics. The filtered signal is passed to a balanced amplitude modulator, M1, whose RF input is supplied from local oscillator LO1.

The figure illustrates the output of the modulator, which consists of a carrier envelope that drops to zero amplitude in the transitions between 0 and 1. The modulated carrier has one phase relation for a 0 and the opposite phase when a 1 is being transmitted. Note the phase of the carrier with respect to the carrier timing tick marks (every 180° of the carrier) shown below the RF envelope. The reduction in amplitude to zero between states is a direct result of the filtering action of LPF1. Were it not for this filtering, the transition between 1 and 0 would not require the amplitude of the carrier to drop, but the occupied bandwidth would theoretically be infinite.

At the receiver, the signal is bandpass filtered in BPF1. From the filter, the signal is supplied to mixer M2, the demodulator. The second input to M2 is from local oscillator LO2. This oscillator must be phase locked to the incoming RF carrier. A number of different carrier recovery circuits can be used. The circuit shown utilizes the fact that the second harmonic of the carrier contains no phase shift (when the frequency is doubled, a 180° phase shift becomes a 360° phase shift, that is, no phase shift at all). The incoming frequency is doubled by the times-two multiplier, ×2, and then used to lock LO2. The output of LO2 is divided by 2 and supplied as the second input to M2.

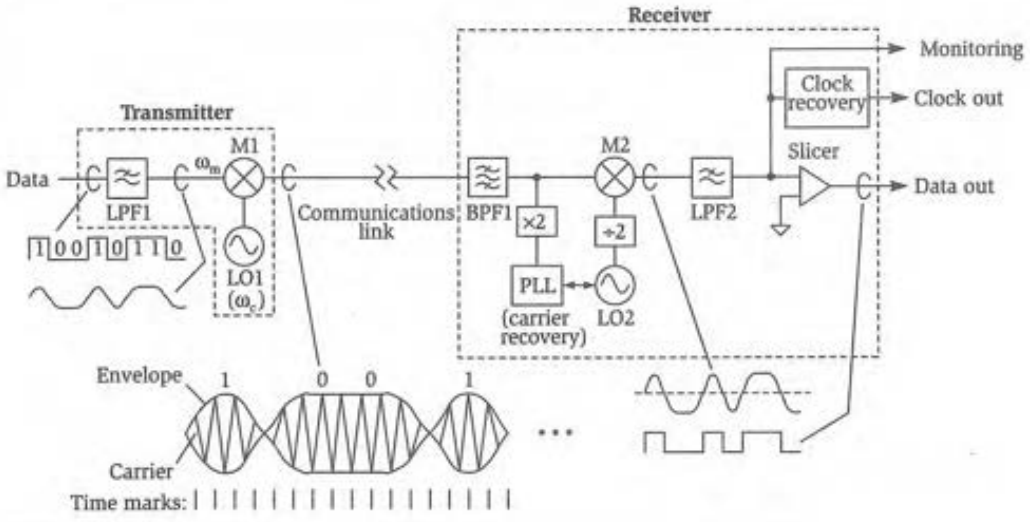


Figure 4.3 BPSK Transmission System.

Basic Math of BPSK

One might be tempted to look at the waveform of Figure 4.3 and conclude that BPSK is a phase modulation technique because the carrier reverses phase between transmission of 1s and 0s. Indeed, the name of the modulation leads to that conclusion. The correct way to visualize BPSK is as double sideband suppressed carrier amplitude modulation.

To understand the modulation process, assume that the data signal is an alternating series of 1s and 0s that has been filtered to a sine wave, following the logic of earlier sections of this chapter. Then the modulating signal can be expressed as

$$e_m = M \sin \omega_m t \tag{4.2}$$

where the subscript m indicates the modulating frequency and $\omega_m = 2\pi f_m$. As implied by Figure 4.3, the modulation process involves multiplying this modulating frequency by the carrier frequency, represented as

$$e_c = A \sin \omega_c t \tag{4.3}$$

The resulting modulated signal is expressed as follows, expanded through use of common trigonometric relations:

$$\begin{aligned} e_r &= e_c \cdot e_m \\ &= AM \sin \omega_c t \sin \omega_m t \\ &= \frac{AM}{2} \cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t \end{aligned} \tag{4.4}$$

The center row shows the phase shift with modulation: When the sinusoid representing the data modulation ($\sin \omega_m t$) is positive, then the carrier has one phase; when it is negative, the carrier phase is reversed, as illustrated in Figure 4.3.

The bottom expression shows that the frequency components of the modulated signal exist at the carrier frequency ω_c minus the modulating frequency ω_m and at the sum frequency. Thus, the bandwidth occupied by the signal is twice the modulating frequency, as is the case with conventional double sideband amplitude modulation. A component at the carrier frequency does not exist.

The mathematics of the synchronous detector shown in Figure 4.3 consists of multiplying Equation (4.3) by the carrier sinusoid and applying more common trigonometric relationships:

$$
\begin{aligned}
e_d &= e_c \cdot e_r \\
&= \sin \omega_c t \cdot AM \sin \omega_c t \sin \omega_m t \\
&= AM \sin^2 \omega_c t \sin \omega_m t \\
&= \frac{AM}{2}(1 - \cos 2\omega_c t) \sin \omega_m t \\
&= \frac{AM}{2} \sin \omega_m t - \frac{AM}{2} \cos 2\omega_c t \sin \omega_m t
\end{aligned} \tag{4.5}
$$

The first term in the bottom line represents the recovered modulation; the second term represents components surrounding twice the carrier frequency. This latter term is filtered in LPF2 of Figure 4.3.

4.2.6 Quadrature Phase Shift Keying (QPSK)

BPSK modulation provides very good immunity to noise and is simple and low in cost. The problem is that it is inefficient in its use of the spectrum — better than FSK, but not good enough for high-speed data transmission. In order to make better use of the spectrum, it is necessary to take some small steps back toward (but not reaching) the analog domain, in that possible states are represented by different analog levels and/or phases. This allows the transmission of more than one bit at a time. The next modulation formats to be discussed are expansions of BPSK, which improve on the spectral efficiency by transmitting more than one bit at a time. The penalties are an increase in complexity and reduced immunity to noise and other transmission impairments.

Figure 4.4 illustrates a basic QPSK (also called 4-QAM for reasons that will become apparent) modulator and demodulator. Data entering the transmitter is split into two channels. For reasons that will soon be apparent, the two are called the I (in phase) and Q (quadrature) channels. Two bits are transmitted simultaneously, one in each channel. Spectral efficiency is improved because the spectrum required to transmit the two bits is no wider than is the spectrum required to transmit one bit using BPSK.

After the data is split into the I and Q channels, each channel is modulated onto a carrier, just as in BPSK. The same carrier frequency is used for each channel, but the phases of the two are 90° apart, or *in quadrature*. It can be shown that the two phases can be transmitted through a common communications channel, ideally without interfering with each other. (In the real world, there are several mechanisms that can cause the two channels to interfere, or *crosstalk*, but crosstalk can be controlled.)

After the two channels are modulated, the modulated carriers are combined for transmission to the receiver. At the receiver, a phase locked loop (PLL) locks a local oscillator, LO2, to the incoming carrier. The signal from LO2 is split into quadrature components. The in-phase component is sent to demodulator M3,

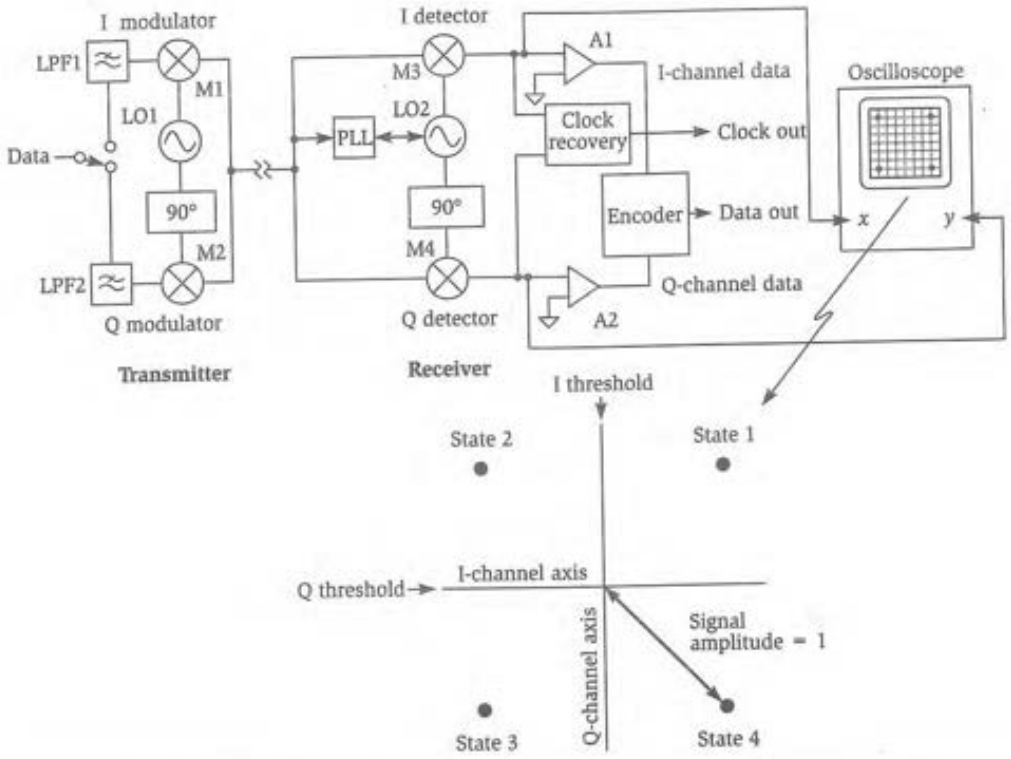


Figure 4.4 Basic QPSK Modulator and Demodulator.

and the quadrature component to the quadrature demodulator, M4. The two quadrature components of the local oscillator demodulate the corresponding component of the combined signal, while the opposite component is cancelled out because it is in quadrature phase relationship. An encoder combines the two bits from the I and Q channels into the resulting datastream. A clock recovery circuit reconstructs the clock signal from the incoming data.

The Constellation Diagram

A particularly useful representation of QPSK and higher-order signals is the constellation diagram, shown at the bottom of Figure 4.4. It is formed by routing the demodulated I signal, before slicing, to the x axis of an oscilloscope and the Q signal to the y axis. Each of the four possible positions of the symbol is formed from the I- and Q-channel data, with the position along the x axis being determined by the I data and that along the y axis by the Q data. The amplitude of the

signal is proportional to the length of the vector from the origin to any one of the four states, which are equidistant from the origin. Note that, in the absence of any transmission impairments, the four points in the constellation are small points. They are enlarged in the illustration to make them easier to see. As the signal transitions from one point in the constellation to another, the trajectory may pass through several other points.

The *y* axis forms the *decision threshold* for the I channel since it is halfway between the I channel states. Similarly, the decision threshold for the Q channel is defined by the *x* axis. The decision threshold defines the signal state at which the comparator(s) for that axis makes the decision that the bit represents one value or the other. If the point in the constellation crosses the decision threshold due to noise or other channel impairment, the comparator will yield the wrong output, resulting in an error in transmission. Computation of bit error rate (BER) is based on determining the probability that any one point in the constellation could be forced to cross a threshold into another state's place in the constellation. When this happens, at least one, and sometimes more, bits are received in error. Obviously, the closer together the states in the constellation (i.e., the "denser" the modulation), the less noise or other disturbance can be tolerated before a state can be misinterpreted as that of a neighbor.

The peak amplitude of the signal is equal to the length of a vector from the origin to the farthest point in the constellation. In the case of QPSK, all four points are equidistant from the origin, so any of them can be taken as the peak amplitude. Note that the constellation diagram represents voltage, not power.

Though the QPSK signal has four states at the same amplitude, between states the amplitude of the carrier drops, due to filtering at the modulator. This filtering is represented by LPF1 and LPF2 of Figure 4.4. Also see Figure 4.3 for an illustration of the resultant modulated waveform. Since QPSK is nothing more than two BPSK signals combined 90° out of phase, then the same waveform of Figure 4.3 would generally apply to a QPSK signal: Between symbols, the signal amplitude passes through 0 if the state of both bits in the symbol changes. This usually doesn't cause much concern in cable television work. However, in the case of systems having high AM/PM conversion, where a change in amplitude of the signal can cause a change in its phase, the amplitude of the signal dropping to 0 can be a problem. AM/PM (spoken as "AM to PM") conversion is a phenomenon in which changes in the amplitude of a signal result in a change to the phase of the signal. In data communications, it can distort the phase reference for the PLL in a receiver, making accurate recovery of the data less likely. AM/PM conversion can be caused by asymmetrical nonflatness of the channel and by capacitance modulation in an amplifier that is operating out of its linear range.

A satellite communications link is susceptible to AM/PM conversion. Often the power amplifier on the satellite is driven slightly into saturation and may exhibit AM/PM conversion. To reduce the effect, satellite communications often

uses *offset QPSK*, in which the transition times of the data in the I and Q channels is offset by one-half of a bit cell. Since the two channels don't change amplitude at the same time, the amplitude tends to be more constant.

4.2.7 Higher-Order QAM Modulation

Binary Numbering

Before we discuss higher-order QAM, we need to review a couple of elementary concepts in binary numbering. We will show groups of two bits whose value is designated as 00, 01, 10, or 11. The correct way to speak this is "zero-zero," "zero-one," and so on ("zero" may be replaced with "oh"). These numbers are not to be confused with the base-10 numbers that look the same. If confusion between base 2 (binary) and base 10 (our familiar numbering system) is possible, then the number will be followed by a subscript 2 or 10. The following are equivalent numbers in base 2 and base 10: $010_2 = 2_{10}$ and $1101_2 = 13_{10}$.

16-QAM Modulation

The next denser modulation format commonly found in cable television work is 16-QAM (16 state quadrature amplitude modulation). It is used for modem return services that require higher bandwidth efficiency than that offered by QPSK. It is similar to QPSK except that each axis is divided into four equidistant levels so that each channel (I and Q) can carry the information of two bits. Each of two phases carries two bits at a time, so there are four bits per symbol.

Figure 4.5 illustrates a basic 16-QAM modulator and demodulator. Note the similarities between this and the QPSK modulator of Figure 4.4. Each channel of the 16-QAM modulator can take on not only the two phases of QPSK but, in addition, intermediate amplitude values. Data is split into two channels, the I and Q channels. Instead of only one bit being routed to each channel at a time, two bits are routed to each. They are added as shown in the table beside the Q-channel modulator: for example, if the two bits are 11 (pronounced "one-one," since we are working in binary numbers), then the output of the adder is +3, and the Q-channel modulator puts out an amplitude of +3, with a positive phase (calling this the "positive" phase is arbitrary). If the data set is 10 (one-zero), then the adder supplies an analog signal of amplitude +1 to the modulator.

Similarly, a 01 data set yields an analog level of −1, which causes the modulator to supply a 1-level signal with the opposite phase. A 00 data set will cause the adder to supply −3 to the modulator, producing a higher-amplitude negative phase signal (saying the phase is "negative" is an arbitrary way to refer to the fact that the phase is opposite that of a "positive" phase). By convention, the inner levels are +1 and −1. To retain equidistance, spacing between levels is 2.

The same thing takes place in the I channel. Each channel can thus take on any of four states. With the I channel in any of four states and the IQ channel

independently in any of four states, a total of 16 states are possible. This is the number of states required to represent 4 bits ($2^4 = 16$). The modulation technique, 16-QAM, derives its name from the 16 possible states of the modulated signal.

The receiver contains in-phase and quadrature demodulators, as does the QPSK receiver. However, the output of each demodulator must be further processed to recover two bits. This is done for the I channel, using the three data slicers shown. The data slicers have thresholds set by the four voltage divider resistors, at levels of T1, T2, and T3. These levels are set halfway between the received states in the I channel. The thresholds are shown in the constellation diagram at the bottom of Figure 4.5. By reading the outputs of the three comparators, A1 through A3, the decoder can deduce the transmitted state. Note that there are more decision thresholds than there were in QPSK, because there are more points in the constellation, and we must have a way to distinguish between all states. We often refer to *decision boxes* rather than

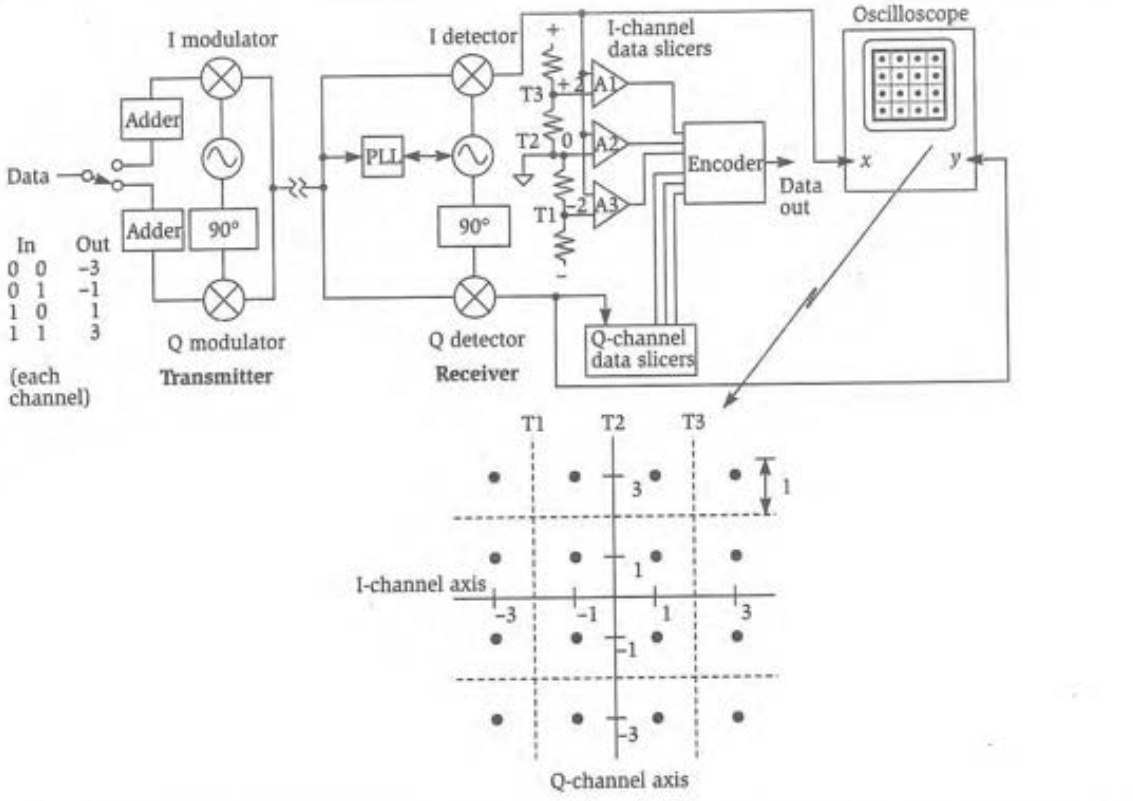


Figure 4.5 16-QAM Modulator and Demodulator.

thresholds, because each point in the constellation has its own box it must be in, defined by the I and Q decision thresholds.

64- and 256-QAM

For downstream transmission, the industry employs modulation formats that are similar to 16-QAM except that they have more states. They transmit more bits per symbol, so the constellation is denser. The distance between states is smaller, so the susceptibility to noise and transmission path distortion is greater. On the other hand, the spectral efficiency is greater, meaning that more bits can be transmitted in the same RF bandwidth.

Figure 4.6 illustrates the constellation diagrams for 4-QAM (i.e., QPSK), 16-QAM, 64-QAM, and 256-QAM. The diagrams are drawn to the same peak scale, so it is easy to see the higher density of the higher-order modulation formats. Also shown for comparison are several characteristics of the different

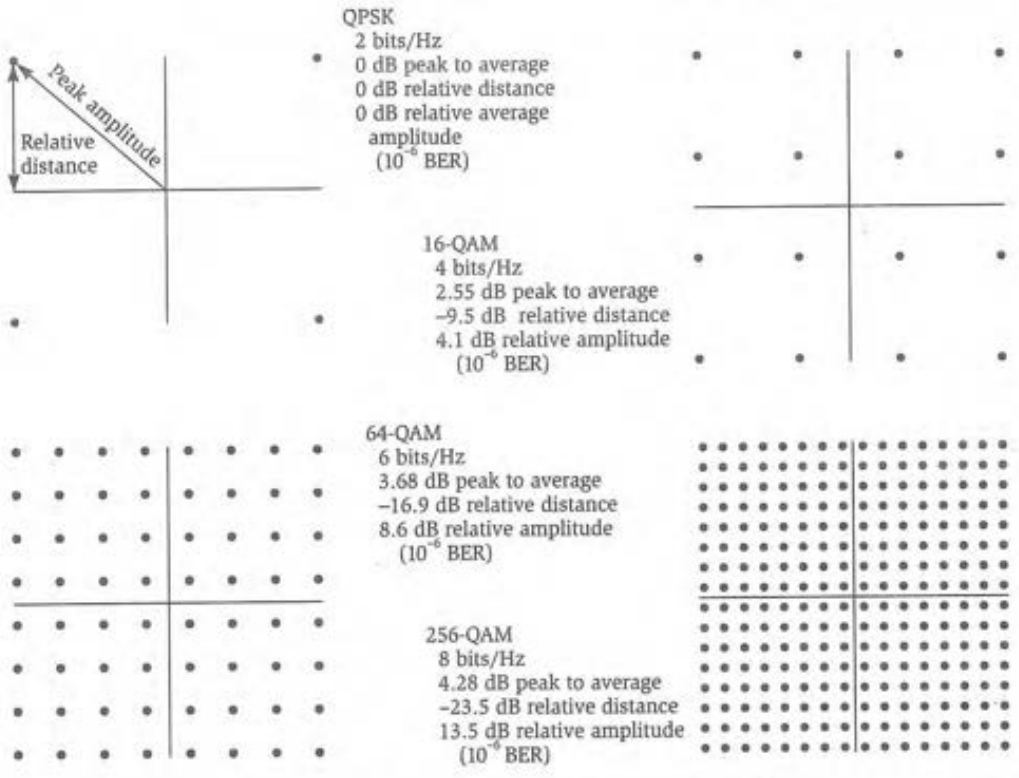


Figure 4.6 Constellation Diagrams, 4-, 16-, 64- and 256-QAM.

modulation formats. Most of the characteristics are shown measured in decibels with respect to a QPSK signal.

For each modulation type, the approximate power peak-to-average ratio is shown first. This is computed by averaging the power of each state, ignoring the effects of filtering. Later we show how this is done. A more precise peak-to-average ratio must include the effects of filtering. Next is shown the theoretical spectral efficiency, in bits per hertz. The spectral efficiency is a figure of merit, defining how many bits per second can be transmitted per hertz of RF bandwidth. The actual spectral efficiency will be less than this, due to the practical need for excess filter bandwidth.

Below the peak-to-average ratio is shown the relative distance between a state and the closest state boundary, expressed in decibels. The relative distance between a state and the boundary is defined in the QPSK constellation, as is the peak amplitude. It is the decibel measurement of how much closer to each other the states are in the denser modulation formats. The reference is a QPSK constellation, and the distance is normalized to the peak, not the average, amplitude.

The relative amplitude entry is derived from the BER vs. signal-to-noise ratio of the equation graphed in upcoming Figure 4.17. It is an indication of the increase in signal-to-noise ratio required to maintain a 10^{-6} bit error rate. Recall that the signal level to which we refer here is the average, not the peak level.

32-QAM Modulation

A similar modulation format to those shown earlier is 32-QAM, which has 32 states and a theoretical spectral efficiency of 5 bits per hertz. Two different 32-QAM constellations are shown in Figure 4.7. Figure 4.7(a) is distinguished from the 64-QAM formats shown earlier in that every second state is occupied. This puts more distance between states, resulting in easier decoding. Figure 4.7(b) illustrates an alternate encoding of 32-QAM, where the four corner states are omitted. Both encoding methodologies have been reported in the literature.

A total of 32 states represents 5 bits ($2^5 = 32$). With 16-, 64-, and 256-QAM, each axis, I or Q, can represent a certain number of bits, because the number of states along each axis is a power of 2. On the other hand, 32-QAM does not have this correspondence between bits and states possible. The five bits making up a symbol can, however, be encoded into one of the 32 possible states.

4.2.8 8-PSK Modulation

Because satellite transponders normally operate close to or in compression, the modulation transmitted over a satellite path must not depend on amplitude discrimination to recover data. For many years, QPSK, or 4-QAM, has been the digital modulation of choice for satellite transmission. Recently 8-state phase shift keying (8-PSK) has begun to see use in order to improve bandwidth efficiency in satellite communications. Figure 4.8 illustrates the constellation

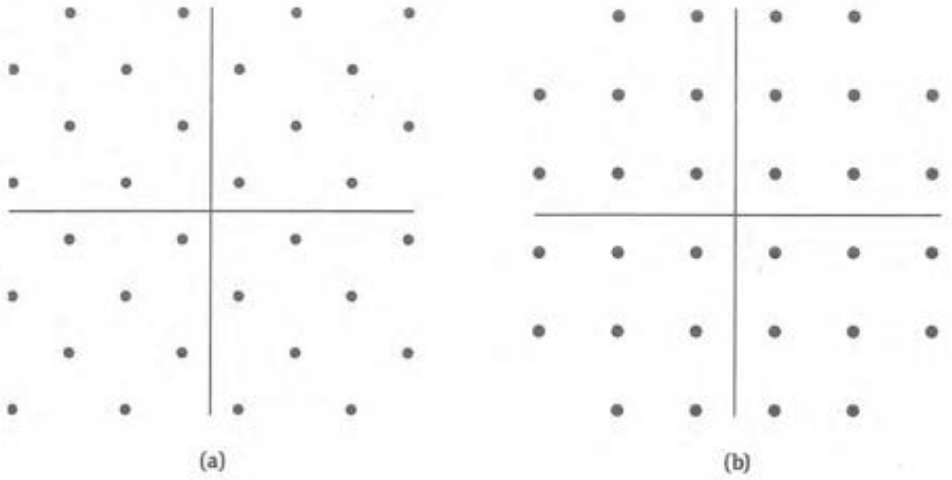


Figure 4.7 32-QAM Constellations. (a) Alternate states. (b) Corners removed.

of an 8-PSK modulated signal. It has eight states (3 bits per symbol); but since they are all equidistant from the origin, all states have the same amplitude (there will be some amplitude reduction between states due to filtering). Thus, the satellite amplifier operates at the same output amplitude at all times, a necessary condition for efficient amplifier operation.

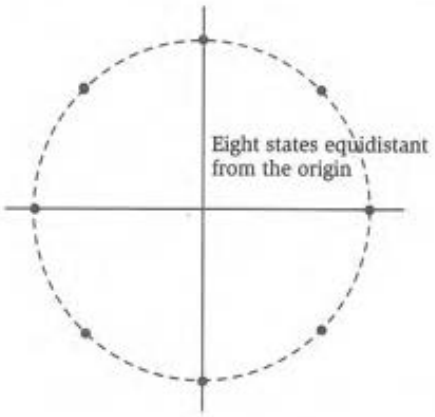


Figure 4.8 8-PSK constellation.

4.2.9 8-VSB Modulation

The FCC has designated 8-VSB modulation as the standard for off-air television in the United States. It is very similar in performance to 64-QAM. In VSB modulation formats, only one phase of the carrier is used, and there is no quadrature component; all bits making up the symbol modulate the same phase of the carrier. That carrier (which is almost completely suppressed) is placed very near the band edge rather than in the center, and one sideband is removed by filtering.

Both 8-VSB and 64-QAM have essentially the same spectral efficiency. 64-QAM carries data on two quadrature phases of the carrier but must place the carrier in the center of the passband. Where carriers in quadrature are employed, it is necessary to transmit both sidebands to allow for separation of the two channels. 8-VSB only uses one phase of the carrier, so separation of components is not an issue. Compensating for the lack of the quadrature phase, 8-VSB is able to transmit twice the symbol rate in the same bandwidth, since the second sideband is not transmitted.

Figure 4.9 illustrates the principle of VSB digital modulation. A set of bits is converted to a parallel word. The number of bits taken at a time is three when 8-VSB ($2^3 = 8$) is being transmitted. These bits are added, and the result is converted to analog and supplied to a balanced modulator. The modulator output is filtered in a vestigial sideband (VSB) filter. Ideally, the VSB filter doesn't pass any of the lower sideband but does pass the carrier. As a practical matter, some of the lower sideband must be transmitted, as in the case of analog transmission. Advances in filter technology since the NTSC television system was developed have made a much smaller vestigial sideband possible. In analog VSB transmission, the picture carrier is 1.25 MHz above the channel edge. In digital VSB transmission, the carrier is only 310 kHz above the channel edge.

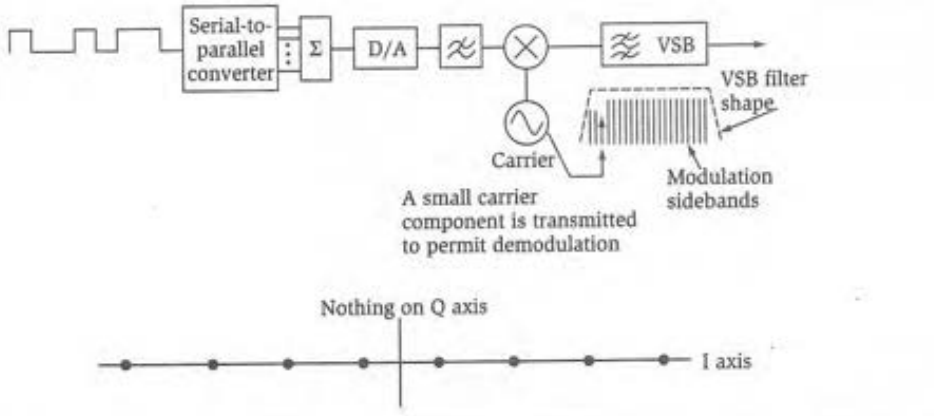


Figure 4.9 VSB Modulator.

A small amount of carrier is inserted in the VSB signal to facilitate carrier recovery. With QAM modulation formats, the carrier may be recovered by taking advantage of the symmetrical sideband nature of the signal. It may be shown that a suppressed carrier can be recovered by doubling the frequency of the received signal, which yields a second harmonic of the carrier. This is true only when the carrier has symmetrical sidebands. Such is not the case with VSB, so a small carrier component must be transmitted to allow recovery.

The susceptibility to transmission impairments of an 8-VSB signal is similar to that of a 64-QAM signal. VSB has a claimed advantage in the presence of phase noise. As shown in upcoming Figure 4.12, phase noise causes the constellation to rotate around the origin. With VSB transmission, the distance to the state boundary on the phase trajectory is much greater than it is with QAM. The excess bandwidth of the broadcast 8-VSB signal is 11.5%. The symbol rate is 10.76 Mbaud, so with three bits per symbol the gross data rate transmitted is 32.28 Mb/s. Due to trellis coding, there are only two information bits per symbol, so the gross payload is two-thirds of this, or 21.52 Mb/s. Adjusting for overhead, the net transport stream payload seen by a transport decoder is 19.393 Mb/s.[1] Sometimes you may see the data rate shown as 19.29 Mb/s. This number is the data rate transmitted (less modulation overhead), where the MPEG sync bit is not transmitted but rather is added back at the demodulator. This data rate (whichever number you use) is less than the 64-QAM data rate transmitted in cable, primarily to allow for more error correction to compensate for the perils of the radio path.

4.2.10 Fundamental Principles of Digital RF Transmission

In this section we describe a number of issues associated with digital modulation. You should be familiar with them in order to make sense of what you experience related to a digitally modulated signal.

Bit Rate and Baud Rate

Perhaps the most common mistake made in communications terminology today is to confuse the rate at which *data* is transmitted with the rate at which *symbols* are transmitted. The rate at which bits are transmitted (the *bit rate*) is measured in *bits per second*. In most modulation formats, however, more than one bit is transmitted at a time. The bits that are transmitted simultaneously are grouped into a *symbol*. The rate at which symbols are transmitted is the *baud rate*.[2]

In DOCSIS upstream transmission, the lowest specified symbol rate is 160 kbaud. If QPSK modulation is used, the bit rate is twice that, or 320 kb/s, because there are two bits transmitted per symbol. However, if 16-QAM is used, the bit rate is four times the baud rate, or 640 kb/s. (Note that the plural of *baud* is *baud*, without an "s.") As shown earlier, 256-QAM modulation groups 8 bits into a symbol, so the baud rate is one-eighth of the bit rate. The

occupied bandwidth of the signal is a function of the baud, or symbol, rate, because the baud rate tells us how many times the modulation state changes each second.

Filtering Methodology

This section introduces the filtering methodology that is followed in many data communications applications and for many types of digital modulation (except FSK, which was covered earlier). The filters at the transmitter and receiver are effectively in cascade in the signal path; together it is desirable that they form a filter just wide enough to convert the highest modulating frequency into a sine wave. (As a practical matter, the filter will be somewhat wider than this.) A particularly useful class of filters is called *raised cosine* filters by the data communications industry. They are called *sine squared* filters by the television engineering community.

The equivalence is seen in the following trigonometric relationship:

$$2\sin^2\alpha = 1 - \cos 2\alpha \tag{4.6}$$

The left side of the equation shows the rationale behind the television usage of *sine squared* terminology, while the right side shows the basis of the data communications term *raised cosine*. The 1 on the right "raises" the cosine term, which by itself ranges from -1 to $+1$. In television work, a sine squared filter is used to produce the well-known sine squared pulse and bar, which are useful in analyzing analog video channel performance. In data communications, the cascaded performance of the low-pass filters at the transmitter and receiver, possibly in addition to the response of the IF filters, form a raised cosine filter, which ensures a fast transition between the modulation states, consistent with minimal spectrum occupancy.

It is desirable to heavily filter at the transmit end in order to reduce the spectrum usage, and heavy filtering is desirable at the receive end to minimize noise added during transmission. However, the total filtering applied must conform to the raised cosine shape. In many communications systems, the designer places one-half of the filter function at each end of the communication channel. When this is done, the filter response at each end is the square root of the complete filtering function, so the filter used on each end is often described as a *root raised cosine* filter. Note that the filters at the transmit and receive ends must be matched to each other and collectively are often referred to as *matched filters*.

Note that any bandpass filter must be symmetrical with respect to the center (carrier) frequency for quadrature signals (that is, any QAM signal). This is a requirement of quadrature transmission systems. If the overall filter response is not symmetrical, we can show that the carrier phase will be recovered incorrectly, resulting in a loss of signal integrity. In addition, the in-phase and quadrature channels will crosstalk.

Required Bandwidth of the Signal

We shall develop the concept of occupied bandwidth using a BPSK signal, and then show how it is extended to any QAM modulation format. The bandwidth requirements of BPSK can be understood by considering the worst case, that of an alternating string of 1s and 0s. This will produce a square wave, one cycle of which is composed of an alternate 1 and 0, or two bits, as explained earlier. Thus, the frequency of the square wave, f_m, is one-half the bit rate. Figure 4.10 illustrates the spectrum of the data signal at baseband. The square wave of alternating 1s and 0s illustrates the highest frequency of the data signal. From Fourier analysis, we know that a square wave consists of all odd harmonics of the fundamental frequency, as shown by the vertical lines in the spectrum plot of the figure.

In real data systems, the data is not an alternating series of 1s and 0s. Rather, it is all possible combinations of strings of 1s and 0s. The resulting spectrum will fill in the spectrum from the maximum frequency, f_m, down to close to 0 frequency (and all harmonics thereof).

However, the harmonics are not needed to convey the information contained in the datastream. If all frequencies above f_m are removed by filtering, then an alternating series of 1s and 0s will produce a sine wave of frequency f_m.

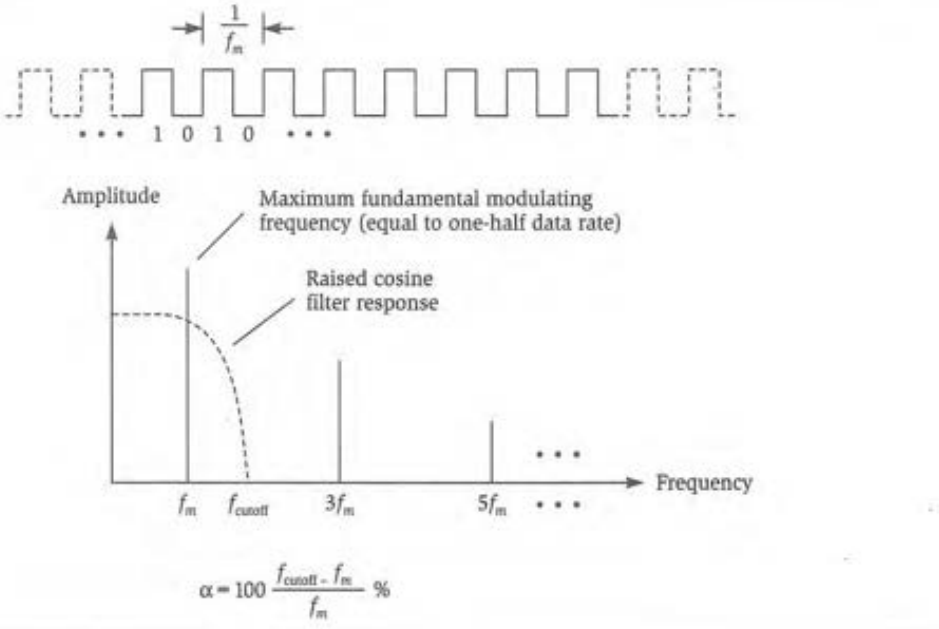


Figure 4.10 Baseband Spectrum of BPSK Signal.

Transitions that occurred at a slower rate (i.e., from a series of like states) would follow a sinusoidal transition path but would be flattened on top.

The *theoretical* (not practical) spectral efficiency of a BPSK signal is developed as follows. The alternating 1–0 sequence defines the highest frequency of the data baseband signal. If it is filtered by removing all harmonics, leaving a sine wave, then the frequency of the sinusoid is one-half the data rate (one cycle is composed of a set of a single 1 followed by a single 0). The bandwidth of the modulated signal is twice the modulating frequency (the sum less the difference between the carrier and the modulating frequency), which in turn is equal to one-half the bit rate. Thus, the spectral efficiency is two times one-half the data rate, or one times the bit rate, one bit per hertz.

A filter that completely cuts off just above f_m may be ideal but is hardly practical. A practical filter has the raised cosine shape and is somewhat wider than f_m. The sharper the filter cutoff (that is, the less bandwidth required between the filter passband and the stopband), the less bandwidth is required. However, the sharper filter is more difficult to build, and more group delay (see Glossary) is likely, depending on the realization. The shape of a real filter is described by the *alpha* of the filter, where alpha is defined as the excess of occupied bandwidth compared with the minimum required bandwidth:

$$\alpha = 100\frac{f_{\text{cutoff}} - f_m}{f_m}\,\% \tag{4.7}$$

Recall that the common filtering practice is to put the square root of the raised cosine filter response at the transmitter, and the same at the receiver. For example, as shown in Chapter 5, the filter specification for the transmitter of a DOCSIS-compliant downstream modulator operating in the 64-QAM mode is an 18% root raised cosine filter, meaning that the filter is 18% wider than the minimum required bandwidth.[3] This may also be interpreted that the occupied bandwidth is 18% wider than the theoretical minimum.

QPSK works the same way as does BPSK as far as the spectrum is concerned. QPSK is simply two BPSK carriers at the same frequency, but 90° out of phase with each other. The spectra overlap; and since a spectrum analyzer doesn't respond to the phase of a signal, the BPSK and QPSK spectra are very hard to tell apart.

Higher orders of QAM modulation are simply the same thing as QPSK, with the exception that rather than each of the two quadrature carriers taking on only two states (phases), they can take on multiple amplitude and phase states. The spectrum is again very difficult to distinguish from that of a BPSK signal. The bandwidth is based on the number of symbols transmitted per second, not on the number of bits transmitted per second. See the earlier discussion of bit rate vs. baud rate.

Clock Recovery

While self-clocking formats, such as Manchester encoding (see Section 4.2.2), could be used with BPSK and higher-order modulation, this is normally not done. The higher number of transitions associated with Manchester encoding would double the required bandwidth of the signal. Rather, an NRZ datastream is transmitted. Clock recovery consists of synchronizing a clock to the transitions in the received data. Where no transition occurs, the clock "freewheels" through the place where the transition should occur. If a long transmission of 0s or 1s occurs, it is possible to lose clock synchronization, so "scrambling" or some other technique is often used to force a state change every so often.

Data Scrambling

Scrambling is often used as a method of dc removal, which is necessary to effect proper operation of a number of circuits in the receiver, including clock recovery. *Scrambling* in its data transmission sense is not the same as when used in the cable television sense. In cable television, the word refers to the process of rendering a video and/or audio signal unintelligible to a normal receiver. A special circuit must be used to restore the signal to its intelligible form.

In data transmission, scrambling is the process of combining a datastream with a pseudorandom bit sequence, with the intent to remove any possible dc component of the datastream. An equivalent way of expressing the function of the scrambler is to say that it equalizes the power density over the occupied bandwidth. It can be shown that, if any data signal is exclusive ORed with a pseudorandom signal having a flat spectrum and no dc component, then the resultant signal will have a flat spectrum and no dc component. The signal can be recovered at the receiver by exclusive ORing the signal with the same pseudorandom signal.

The exclusive OR operation is a logical operation on a single bit of data. If the two bits are the same, the output of the exclusive OR operation is a logical 0; if the two bits are different, the output is a logical 1.

Differential Encoding

The phase locked loop used for carrier recovery in Figure 4.3 exhibits a problem, in that a phase error ambiguity exists: If the output of LO2 divided by 2 is in phase with the carrier applied to M1, then the recovered data will have one polarity; if it is out of phase, the recovered data will have the opposite polarity. Since the loop is closed at twice the frequency of the carrier, no simple way to determine the correct phase exists. One way would be to detect a code word buried in the data and to correct for a phase error.

A method frequently used to resolve the phase ambiguity is *differential encoding*. In differential BPSK, a 0 is represented by no change in phase from

the previous state; a 1 is represented as a 180° phase change. The detector is arranged to compare the phase of the previous bit with the phase of the present bit; if the phase doesn't change, a 0 is delivered as the received bit. If the phase is changed, a 1 is delivered (or vice versa). With more complex modulation schemes, sometimes a recognizable binary word is used to determine the correct phase.

The Eye Diagram

The recovered waveform at the "monitoring" point shown in Figure 4.3 should be the same as the waveform applied to modulator M1. However, distortion in the transmission path will change the wave shape, as suggested in Figure 4.11 (the dotted line is the original wave shape; the solid line is the received wave shape). Noise will put "ripple" on the waveform, as seen on an oscilloscope. If enough noise is added, then the state of the signal may be misinterpreted, producing a bit error. Besides noise, group delay can distort the wave shape, as can echoes and amplitude tilt in the channel.

A particularly useful oscilloscope display is called an *eye diagram*, so named for its resemblance to an eye. It is produced by supplying the received

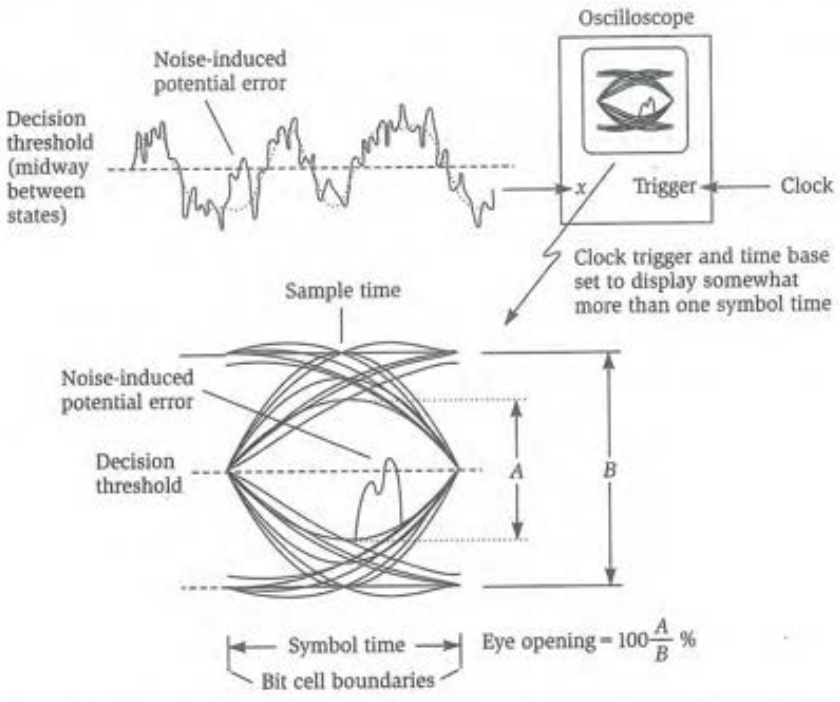


Figure 4.11 BPSK Received Waveform and Eye Diagram.

waveform (before the slicer) to the vertical input of the oscilloscope. The horizontal time base is set to display more than one bit cell and is triggered by the recovered clock. (For simplicity, the diagram shows only one bit cell width, though typically some additional time would be displayed.) When a bit is undergoing a transition from 0 to 1 at the trigger time, the waveform will be positive-going on the left. When it is in transition from 1 to 0, it is negative-going. It may also be the same as the last bit, in which case it forms a flat line at the top or bottom of the display. Similarly, the display toward the right is formed by transitions to the next bit after one displayed.

The slicer interprets the signal to be either a 1 or a 0. The sample time should be in the center of the bit cell, when the eye is as "open" as possible. One of the practical error sources, which is often not considered in computing BER versus noise, is errors in the timing of the sample: If the timing is in error because of problems in the clock recovery circuit (perhaps due to noise, interference, or waveform distortion), then it is more probable that the bit will be misinterpreted. A figure of merit is the *eye opening*, or the percentage of the peak (fully opened) eye that is seen as open in the eye diagram.

Eye closure can be caused by a number of transmission problems. The effect of noise is illustrated in Figure 4.11. Besides noise, errors in the filter shape can cause pulse distortion, which will tend to close the eye. If the transmission path exhibits excessive group delay or if an echo exists, then one bit may not completely "go away" before the next bit starts. This condition is called *intersymbol interference*, or ISI.

Illustrated in Figure 4.11 is a potential bit error, in which a noise spike carries the waveform so far from its correct position that it actually crosses the decision threshold of the slicer. If the clock causes the slicer output to be read at the time of the noise spike, then a bit error would occur.

We have illustrated the eye diagram for a BPSK signal. When quadrature modulation is used, such as in 4-, 16-, 64-, or 256-QAM modulation, the eye diagram is displayed separately for each axis of modulation. When there are more than two states per symbol, then the eye diagram will be more complex but will nonetheless resemble that of Figure 4.11.

Effects of Interference on the Constellation Diagram

The constellation diagram is useful for understanding the effect of interference on the signal. In Figure 4.12, quadrant I illustrates the effect of a continuous wave (CW) interfering signal in the passband of the receiver. An interfering carrier creates a beat note, which forces each point in the constellation to spread out in a circle around the desired point. The radius of the circle is proportional to the relative amplitude of the interfering carrier. Unlike analog video signals, the impact of the interfering carrier tends to be independent of its location within the passband, except for some effect caused by nonflatness of the receive filter.

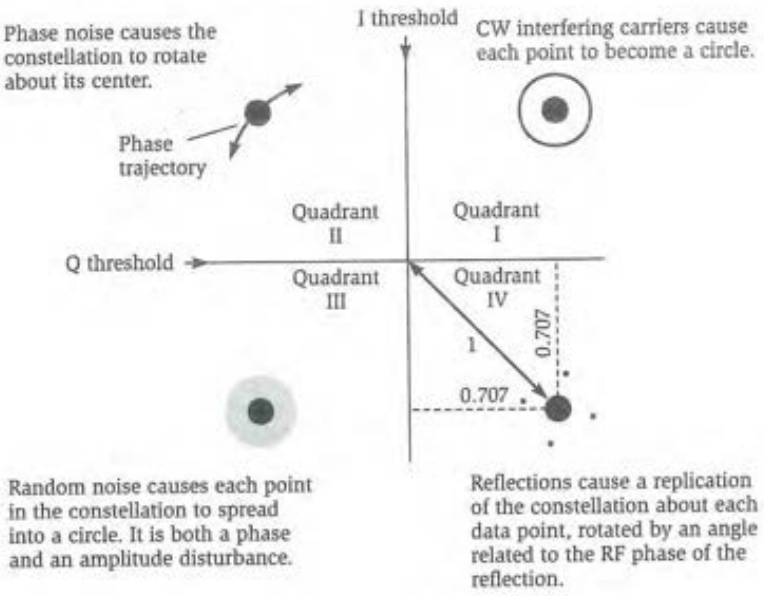


Figure 4.12 The Effect of Interference on the Constellation Diagram.

It is worth noting that composite second order (CSO) and composite triple beat (CTB) are not CW interferences. Their amplitudes change constantly, with peaks reaching 15 dB above the average. The variation is due to phase alignment of the multiplicity of carriers. The occurrence of a peak is rare, but the peak can last from a microsecond to a few hundreds of microseconds (longer than the burst protection period of the Interleave.[4]

Quadrant II illustrates the effect of phase noise on the signal. Phase noise is introduced when a local oscillator has significant phase noise. It can also be shown to occur when the transmission channel does not have a response that is symmetrical about the carrier. Phase noise causes the constellation to rotate about the origin.

Random noise (quadrant III) causes each point in the constellation to "blur," somewhat as with CW interference, except that the spreading of the point tends to be more nearly uniform within the circle, and the circle does not have a well-defined radius. Again, recall that the spreading of points in the constellation becomes a problem when the spread points cross a threshold to another state. With QPSK, that threshold is defined by the axes of the constellation diagram. With more dense modulation, the thresholds are closer together, so lower levels of interference will cause the threshold to be crossed. As a practical matter, an error may occur when a point is forced close to a threshold boundary: Real demodulators don't work quite as well as the theory predicts.

Finally, reflections, which cause ghosting in analog transmission, cause a replication of the constellation around each point in the constellation, as illustrated in quadrant IV. The constellation is rotated with respect to the "direct path" constellation, due to phase shift between the direct and reflected RF carriers. Usually, systems that employ denser modulation formats include adaptive equalizers, which can compensate for echoes if the echo doesn't change too quickly.

If the peak amplitude of the signal is 1, as shown in quadrant IV, then the distance from the nominal location of any of the four states to the nearest state boundary is 0.707. This result is easily obtained via simple geometry, where the amplitude vector, of length 1, is the hypotenuse of a right triangle and the distance to the nearest decision threshold is one of the two other sides of the triangle. Bear in mind that it is more common to measure the amplitude of a digital signal by using the average amplitude, which concept is explored later.

Effect of Gain Compression on the Constellation Diagram

Gain compression can occur in any amplifier stage, but it is most likely to be recognized as such if the amplifier is handling only a single digitally modulated signal. Gain compression may be described as a voltage transfer function having second- and third-order components, just as when we are describing distortion in broadband amplifiers. As the input signal instantaneous amplitude increases, the output amplitude increases more slowly. If an amplifier is handling multiple signals, this phenomenon produces composite second- and third-order distortion.

Figure 4.13 illustrates the effect of gain compression on the digitally modulated signal. The uncompressed constellation points are represented as dots, and the compressed constellation points are represented as pluses. In both cases we have normalized the signal amplitude to the same average power. This represents what a receiver would do if its automatic gain control (AGC) responded to average power. The decision thresholds are at the same point for both the uncompressed and the compressed constellations, due to the normalization. Note that the uncompressed constellation points are in the center of the decision boxes, as far from the decision boundaries as possible, where they should be for best data recovery.

The compressed constellation has very many points near the edge of the decision boundary; a slight amount of noise or other impairment would drive the constellation point outside of the boundary, forcing a bad decision on the transmitted information. For illustration, we have chosen a rather large amount of gain compression. But you can see that the problem gain compression causes is that the constellation is moved from the optimum location, with the result that more errors will be made for a given impairment level.

Peak-to-Average Ratio

Higher-order modulation techniques exhibit a higher peak level than the average level read on a power meter. Figure 4.14 illustrates one reason for this, based on

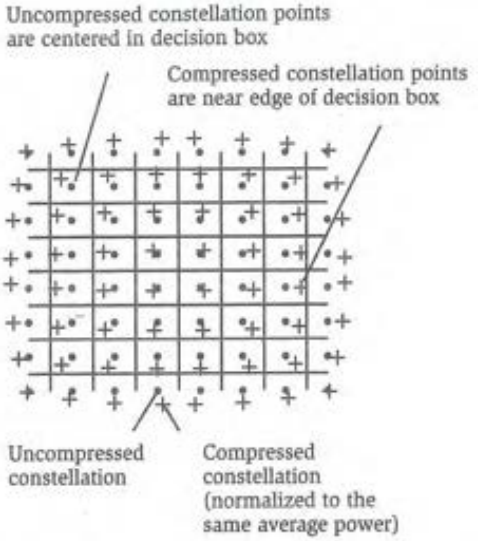
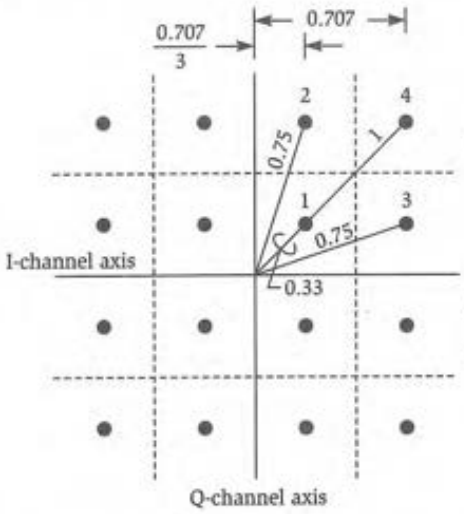


Figure 4.13 Effect of gain compression on a 64-QAM signal.



Element	Length (voltage)	Length^2 (power)
1	0.33	0.11
2	0.75	0.56
3	0.75	0.56
4	1.00	1.00
Average:	0.71	0.56

Peak to average: $20 \log \frac{1}{.71} = 3$ dB $\quad 10 \log \frac{1}{.56} = 2.55$ dB

Figure 4.14 Peak and Average Level of a 16-QAM Signal.

a 16-QAM signal. It shows the same constellation diagram as depicted in Figure 4.5. The constellation is scaled such that the amplitude of any one of the four corner states is 1. This represents the peak level of the signal. The chart in the figure computes the average signal level based on equal occupancy of any of the 16 possible states. Since all four quadrants behave the same way, it is necessary to average over only one quadrant.

The voltage amplitude of each state is determined by simple geometry, assuming that the highest-amplitude state has an amplitude of 1. From the geometry of a right triangle, it then follows then that the coordinates of that state are (0.707, 0.707). The state closest to the center has a distance on each axis of one-third of this, so its total distance is one-third that of the farthest state. States 2 and 3 each have one coordinate of 0.707 and one coordinate that is one-third of that. Application of the Pythagorean theorem shows that the magnitude is

$$\sqrt{0.707^2 + (0.707/3)^2} = 0.75$$

When looked at on a voltage basis, the voltage average is the average length of the four vectors that define the states in the quadrant. The power average is obtained by squaring the voltage vector for each state, because power is equal to the square of the voltage divided by the resistance:

$$P = \frac{E^2}{R}$$

The resistance would cancel out in the computation, so it is omitted. The right-hand column in the chart in Figure 4.14 computes the power peak-to-average ratio, based on squaring the voltage vectors.

Effects of Fitering on Peak-to-Average Ratio

The peak-to-average ratio is not completely characterized by the state distribution as implied, however. It is also a function of the filtering at the transmitter. This was illustrated in Figure 4.3, which shows an amplitude dip in a BPSK signal when the transmitted bit changes states. The same thing happens with higher-order modulation techniques, and that somewhat changes the peak-to-average ratio. The peak-to-average ratio is of concern in cable systems, particularly in the operation of return plant, where a limited number of digital signals comprises the return spectrum handled by the return laser.

It was shown that the peak-to-average ratio is partially a function of the relative amplitude of each state and the percentage of time the signal spends in each of those states. Another factor, not yet taken into account, is the filtering applied to the signal. As the filtering becomes more aggressive (i.e., lower alpha, corresponding to smaller excess RF bandwidth), the filter tends to ring more,

meaning that the signal will be of even greater amplitude than predicted, for a small percentage of the time.

Ito[5] studied this theoretically for a filter having an alpha of 11.52%. He computed the percentage of time the power of a 64-QAM signal exceeded its average by various amounts. The curve follows the percentage computed based on state occupancy, until the percentage of time is less than the percentage of time the signal spends in the highest state. He found that the signal exceeded the average by 6 dB or more for about 0.1% of the time and was 9 dB above the average for about 0.00001% of the time. (For reference, 6.25% of the time a 64-QAM signal is in its highest-amplitude state, which is 3.68 dB higher than the average power.)

Multiple-carrier systems, such as OFDM (described in Section 4.3.3), had a much higher peak-to-average ratio. It exceeded the average by 12 dB for 0.00001% of the time. This is because multiple-carrier modulation signals tend to have the amplitude distribution of random noise, which has a finite probability of being very high for a very short time. The subject of peak-to-average ratio is revisited in Chapter 16.

Effect of Linear Distortion on Pulse Shape

Linear distortions that can enter the cable plant include *echoes*, also called *reflections*, in which a pulse is reflected from an impedance mismatch and adds constructively or destructively to the original waveform. Another linear distortion is *group delay*, in which the delay of the channel is not flat with frequency. Yet another linear distortion is *amplitude ripple*, in which the frequency response of the channel is not flat. All of these linear distortions are familiar to those who have dealt with them in analog modulation. In digital modulation, these distortions tend to close the eye of the data (illustrated in Figure 4.11) and can present a serious impediment to data recovery.

Figure 4.15 illustrates pulse distortion caused by a single reflection. In the modeling, the reflection was delayed about 75% of the pulse duration. In many cases the pulse would be delayed much more. Also, to make it easier to see the effect of the delayed pulse, we have assumed a much wider excess bandwidth than is normal. You can see that the pulse is significantly distorted, which would result in significant closure of the eye diagram. Since the delay in the reflected signal may be much more than one symbol time, we often refer to these distortions as producing *intersymbol interference*, or ISI.

Adaptive Equalization

Because ISI can be so serious a problem in dense modulation formats, it is desirable to compensate for it at the receiver. This is done with an *adaptive equalizer*, a circuit that produces an echo that is the same as the echo added by the system, except the intentionally produced echo is subtracted from the signal.

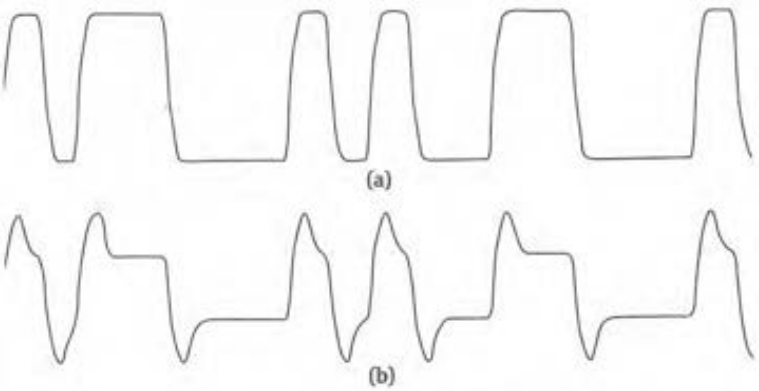


Figure 4.15 Pulse distortion caused by a reflection. (a) Original pulse. (b) Distorted pulse.

The higher the density of the modulation (e.g., 256-QAM), the more important the adaptive equalizer is, because less eye closure can be tolerated before bit errors occur.

Figure 4.16 illustrates an adaptive equalizer installed in each channel of a QAM demodulator. Figure 4.16(a) shows a QAM demodulator, with an adaptive equalizer installed in each channel, after the demodulator and before the data slicers. Thus, we have an analog signal at the point where the adaptive equalizers work. (Note that in modern circuitry, the signal is often digitized before detection, and the detection, filtering, and equalization are performed using digital signal-processing techniques. The point is that the signal at the adaptive equalizer is a continuous signal, even if it is represented digitally. It has not been reduced to the final binary value, which is done in the data slicers and the encoder.)

Figure 4.16(b) shows a 24-stage adaptive equalizer. The signal from the demodulator enters at the left and is successively delayed by small intervals (usually one symbol in time), designated Z^{-1}. After each delay, the signal amplitude is adjusted by complex gains A1, A2, etc. The signal with gain adjusted is added to the main signal, which is shown coming through A1 before any delay, plus the output of all other stages of the adaptive equalizer. Note that the gains A may be negative values, to subtract signal if necessary; depending on the RF phase of an echo, the reflected signal may add, subtract, or have any phase angle in between.

Thus, the adaptive equalizer generates an echo similar to that introduced in the plant and then subtracts it from the main signal in order to cancel the echo. The small eye diagrams at the left and right illustrate the effect of the adaptive equalizer (for simplicity, we are showing a BPSK eye, but the adaptive equalizer is more likely to be needed for dense modulation). Of course, it is necessary to adjust the value of each "tap" (each gain A) to compensate for whatever deficiency exists. The gains A may be positive or negative. It is the job of the

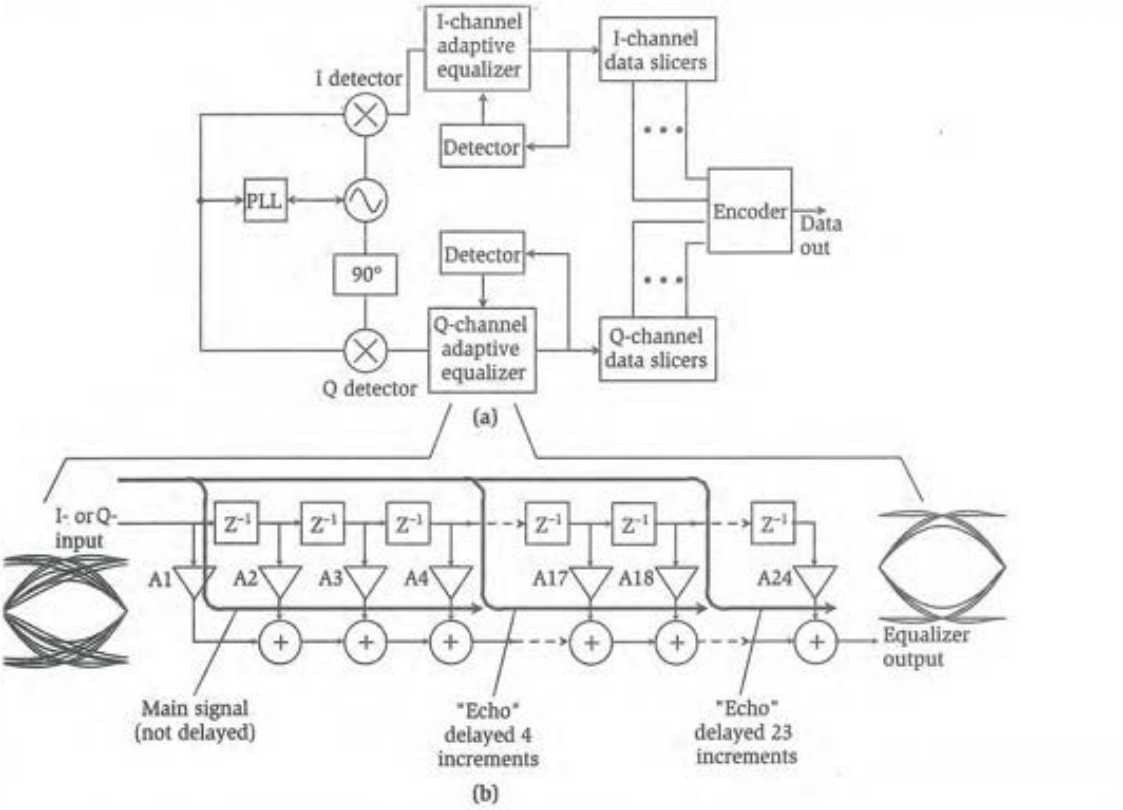


Figure 4.16 (a) Location of adaptive equalizer in receiver. (b) Adaptive equalizer.

detector, shown by each adaptive equalizer, to determine the eye closure and to adjust the taps until the eye diagram is optimized.

Adaptive equalizers are useful for compensating for echo and also for compensating for group delay and amplitude response. In some instances the main tap will not be the first one, because preechoes are possible. Amplitude errors in the channel may be modeled as a pair of echoes, one before and one after the main signal, so the detector may have to assign another tap the job of passing the highest signal (highest value of A), in order to have some preecho taps to use in canceling amplitude errors. A common combination is 7 preecho, the main tap at the eighth position, and 24 posttaps.

In the past it has been most common to place the adaptive equalizer in the receiver, as is done in digital video receivers. However, DOCSIS modems have a unique need in the upstream direction that has caused the equalizer to be placed in the upstream transmitter (at the modem) and the detector in the Cable modem

termination system (CMTS). Since signals from each modem travel a partially different path, they are subject to different distortions, depending on that path. Thus, each modem will require a different adaptive equalizer setting. A detector in the CMTS monitors the eye opening for each modem and sends signals to that modem telling it to adjust its adaptive equalizer filter until the eye opening is maximized. After that has been done, the eye opening is checked on each transmission, and changes are commanded from the CMTS to the modem when necessary.

4.2.11 Defining the Level of a Digitally Modulated Signal

Analog television signals are measured at the peak amplitude of the signal, which corresponds to the sync tips. This level is constant with modulation. The formal definition of the signal level is "the RMS value of the carrier amplitude during sync tips."

Digital signals are sometimes measured based on peak level, but it is standard in cable TV practice to measure the signal based on average level. A thermocouple power meter, which measures average power, is the basic instrument used. Measurement of the average level is more accurate, because it can be measured by supplying the signal to a thermocouple-type RF power meter. The power read on the power meter is usually expressed in decibels with respect to 1 milliwatt (dBm). Such meters are simple and quite accurate. As normally used, a spectrum analyzer comes close to measuring the average signal level (after corrections), not the peak.

The formal interpretation of "signal level" as applied to digital signals is as follows.

> The definition of the level of a digital signal shall be the power level as measured by a power meter which uses a thermocouple as a transducer. That is, the measurement shall be the average power in the signal, integrated over the actual occupied bandwidth of that signal. The signal level should be presented in decibels with respect to 1 millivolt RMS in a 75-ohm system. Thus the measurement reported is the RMS value of the sinusoidal voltage that would produce the same heating in a 75-ohm resistor as does the actual signal.[6]

This definition simply means that the reported signal level is the level of a CW signal that would produce the same heating in a resistor as does the actual signal. The signal level is converted from dBm (commonly reported by power meters) to the equivalent value in decibels with respect to 1 millivolt in a 75-ohm system (dBmV), the common unit of measure of signal strength in cable TV practice. The relationship between signal level measured in dBm and that measured in dBmV, assuming a 75-ohm system, is the same as in the analog domain:

$$\text{signal level (dBmV)} = \text{signal level (dBm)} + 48.75$$

4.2.12 Bit Error Rate Versus Carrier-to-Noise Ratio

It is appropriate to discuss the bit error rate versus the carrier-to-noise ratio of the signal. Usually in data communications, though, the independent variable is not taken to be carrier-to-noise ratio but is an equivalent measurement, expressed in terms of the energy per bit (E_b) divided by the noise per hertz of bandwidth (N_0). We explain this concept and then discuss the requirements for various modulation formats.

The Concept of E_b/N_0

For a given modulation format, as one increases the data rate, the required transmission bandwidth will increase proportionally. If a 1-Mb/s datastream requires 0.5 MHz of transmission bandwidth, then a 2-Mb/s datastream will require 1 MHz of bandwidth. The bandwidth required has doubled, doubling the noise that of necessity must be admitted by the filtering, but the data rate has also doubled.

In order to compare the performance of different modulation formats, which may be used to transmit different bit rates, we use the concept of energy per bit divided by noise power per hertz, E_b/N_0. This parameter, pronounced "E B over N zero," removes data rate and noise bandwidth from the expression. It serves the same purpose as a carrier-to-noise ratio in analog video transmission, but it is not the same as carrier-to-noise ratio. One can think of this parameter as the average carrier-to-noise ratio per bit. In a given case, one can convert to carrier-to-noise ratio by multiplying E_b/N_0 by the bit rate and then dividing by the noise bandwidth. In logarithms, this is done as follows:

$$\frac{C}{N} = \frac{E_b}{N_0} + 10\log(\text{bit rate in b/s}) - 10\log(\text{noise bandwidth in Hz})$$

The use of E_b/N_0 permits one to compare the performance of different modulation systems without having to correct for different bit rates between the systems.

BER Versus E_b/N_0

The bit error rate of various levels of QAM modulation can be approximated by the equation[7]

$$P_M = 2\left(1 - \frac{1}{\sqrt{M}}\right) \operatorname{erfc}\left(\sqrt{\frac{3k\gamma_b}{2(M-1)}}\right)\left[1 - \frac{1}{2}\left(1 - \frac{1}{\sqrt{M}}\right) \operatorname{erfc}\left(\sqrt{\frac{3k\gamma_b}{2(M-1)}}\right)\right],$$

where

M = number of states per symbol (e.g., 64 for 64QAM)
γ_b = average C/N ratio per bit (E_b/N_0)

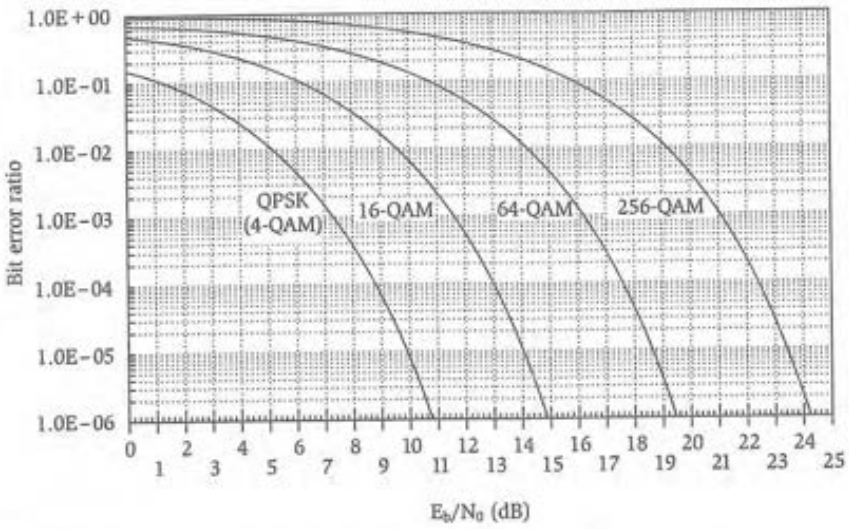


Figure 4.17 Bit Error Ratio vs. E_b/N_0 for Several Levels of QAM.

The error function, erfc, is a mathematical expression of the probability that a normal random variable (such as random noise) will exceed a certain level.

Graphing this expression for a number of levels of QAM yields the chart of Figure 4.17. This graph is normalized for a carrier-to-noise ratio, expressed as E_b/N_0, the energy per bit divided by the noise per hertz of bandwidth. This is a common expression for carrier-to-noise ratio, which is independent of bit rate and bandwidth. This graph does not include the effects of other impairments from the signal source, the transmission channel, or the receiver.

Note that digital communications engineers often refer to this as *signal*-to-noise ratio rather than *carrier*-to-noise ratio. Cable television engineers often reserve the term signal-to-noise ratio, for the baseband measurement and use carrier-to-noise ratio for the RF measurement. Digital communications engineers don't always make this distinction.

4.3 Forms of Spectrum Sharing

This section discusses several forms of spectrum sharing, in which multiple users share the same spectrum.[8] The basic reason for spectrum sharing is better efficiency of a scarce resource (spectrum). Some of the forms shown also claim advantages in the efficacy of data transmission. Note that these forms do *not* represent new modulation techniques — they generally use some level of QAM

modulation — but they do divide the spectrum differently. Nonetheless, some writers have used the term *modulation* to describe them. The careful reader will note similarities among some of the techniques. Differences may sometimes be subtle.

4.3.1 Time Division Multiple Access (TDMA)

One mature method of sharing spectrum is time division multiple access (TDMA), which is used extensively in data systems, thanks to its relative simplicity and maturity. It is quite useful for return path transmission, in which a number of return transmitters need to communicate to the headend.

Figure 4.18 illustrates time division multiple access. The reference system is a number of transmitters in one node, each of which must communicate back to the headend. In TDMA systems, each transmitter transmits on the same frequency but at different times. After each transmitter has its turn, the sequence begins again. One set of transmissions from each transmitter is called a *frame*. Ideally the level of each transmitter is adjusted such that a consistent receive level is measured at the headend. TDMA systems are covered in more detail in Chapter 6.

Some guard time must be allowed between individual transmissions, to allow for errors in the transmission time of each individual transmitter. The receiver must also be able to lock to each transmission when it starts. Of course, since each transmitter is not transmitting full time, the transmission rate must be fast enough to allow for each transmitter to transmit its complete set of bits during its turn. The bandwidth required is increased accordingly. A synchronizing signal must be sent from the headend, which defines the time slots.

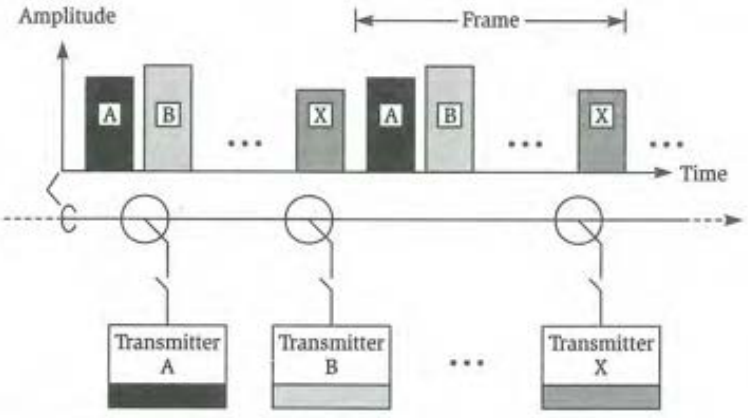


Figure 4.18 Time Division Multiple Access.

It is not necessary that each time slot (A, B, and X in Figure 4.18) be of equal length. Several protocols, including DOCSIS, allow the length to be varied, depending on the amount of data to be transmitted. This is explained in Chapter 5.

For downstream signal transmission, where only one transmitter is involved, a similar system can be used. The only difference is that it is not necessary to turn off one transmitter and turn on another. This transmission method is called *time division multiplexing*, or TDM. It is covered in more detail in Chapter 6.

4.3.2 Frequency Division Multiplexing (FDM)

Another long-established method of sharing spectrum is frequency division multiplexing (FDM), as illustrated in Figure 4.19. FDM systems use a separate narrow subchannel for each transmission, dedicating the channel to that transmitter. Compared with TDMA systems, FDM requires multiple receivers at the headend, though arguably digital signal processing (DSP) techniques significantly ease that burden. (DSP uses a specialized microprocessor to implement signal-handling functions digitally. Modulation and demodulation functions, for example, can be described mathematically. Traditional circuitry approximates the appropriate mathematical functions. DSP techniques directly implement the appropriate mathematics.)

A guard band is required between channels, to allow for practical filtering limitations. This causes some loss of efficiency, though arguably TDMA systems also exhibit some loss of efficiency due to the need for guard times between transmissions. Any suitable modulation method may be used by each transmitter. It is possible to monitor the error rate on each and to vary the modulation format to suit the quality of the channel.

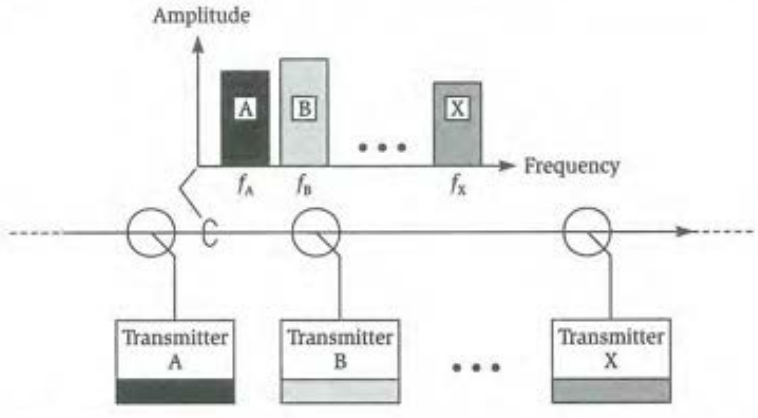


Figure 4.19 Frequency Division Multiplex.

4.3.3 Orthogonal Frequency Division Multiplexing (OFDM)

Figure 4.20 illustrates the idea of OFDM transmission. The incoming data is broken into a number of individual datastreams, each of which is then modulated onto its own carrier. The carriers are spaced just far enough apart to prevent undue adjacent channel interference. Each carrier has modulation sidebands, as illustrated by the dashed lines in the figure (separated for clarity). The carriers may be moved closer together than one would normally expect based on the data rate if the timing of bits on adjacent carriers is carefully controlled. The adjacent signals are then said to be *orthogonal*.

Any convenient modulation may be used on each of the carriers. QAM is implied in the figure. To generalize the type of modulation, communications engineers often use the term *m-ary* to refer to a modulation of any type. Note that OFDM is not itself a modulation technique, though you will often see it described as if it were.

By dividing the spectrum as shown, one replaces a single band of high-speed data with a large number of bands of lower-speed data. Since demodulation is done over a narrower band, the effects of impulse noise and group delay are reduced, and the effective carrier-to-noise ratio is improved in each channel by the narrowing of the passband.

It is possible to monitor the error rate on each carrier individually. If one channel develops an unacceptable error rate, its density of modulation can be reduced (e.g., go from 16-QAM to QPSK modulation). This will necessitate reducing the bit rate on the carrier, meaning that either some bits will have to be reassigned to another channel or the overall bit rate will have to be reduced.

For return path services, it is possible to allow bits on different subchannels to originate from different transmitters. In this case, the technique has been called orthogonal frequency division multiple access (OFDMA). It is also possible to utilize each subchannel in a time division multiple access form, where

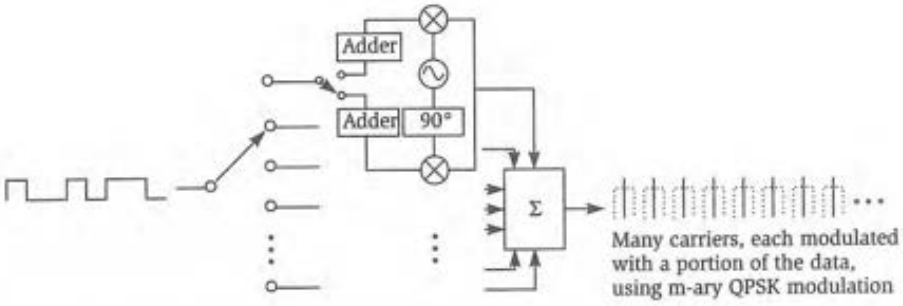


Figure 4.20 Principle of OFDM Transmission.

each subchannel is shared by many transmitters, one at a time. This has been called variable constellation multitone (VCMT).

A closely allied sharing method is synchronous discrete multitone (S-DMT). Adjacent subchannels are synchronized such that spectrum overlap is minimized, allowing closer spacing of carriers. This effectively negates the disadvantage of having to allow guard bands between channels. It does require careful synchronization between transmitters, however. Adjacent channels so synchronized are said to be *orthogonal*. Yet another similar method is known as discrete wavelet multitone (DWMT). The method again assigns small subchannels to carry data in long symbols (i.e., a low data rate in each subchannel). You might observe that some of these methods are different names for the same basic technique.

4.3.4 Code Division Multiple Access (CDMA)

Code division multiple access (CDMA) is a form of spread spectrum communications that is being used for, among other applications, some second-generation cellular phone systems (others use TDMA). The method is also known as S-CDMA, where the S stands for *synchronous*. S-CDMA is one of the advanced upstream options specified for DOCSIS 2.0.

Figure 4.21 illustrates the principle of CDMA transmission. The basic modulation is often QAM of some level. The data to be modulated on each axis is exclusive ORed with a higher-speed pseudorandom bit sequence (PRBS), which is also known as a *spreading code* or *spreading sequence*. The signal is then modulated, normally using m-ary QAM. The effect is to spread out the carrier over a wider bandwidth than necessary for carriage of the information.

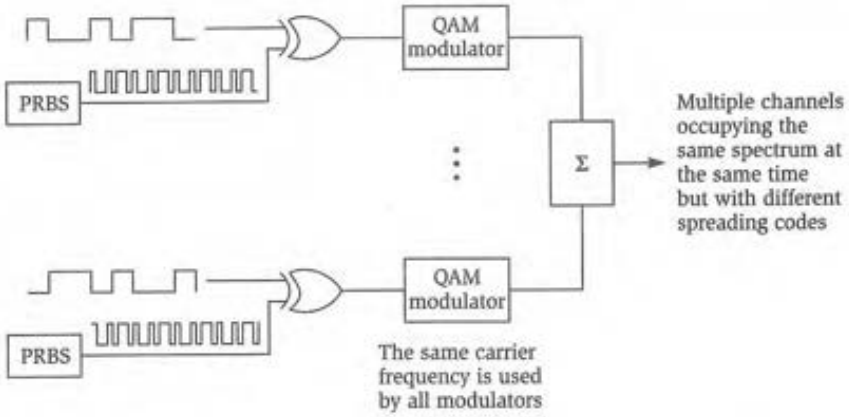


Figure 4.21 Principle of CDMA Transmission.

At the receiver, the demodulated signal is again exclusive ORed with the same PRBS. The result is a collapse of the transmission bandwidth back to that required to convey the information signal without the PRBS. Recovery of the original signal can then proceed. In the process, any narrowband interference that is added to the signal during transmission is spread, most of it out of the receive channel. In addition, if an impairment, such as a suck-out, affects part of the channel, the signal quality is minimally degraded, because most of the energy lies outside of the affected portion of the passband. (For the benefit of data communications engineers, the term *suck-out* is used in the cable television industry to mean a narrow frequency-selective reduction in the response of a transmission path.)

Figure 4.21 illustrates that many intentional transmissions may occupy the same frequency band simultaneously without interfering with one another, so long as each uses a different spreading sequence. All signals may be received simultaneously, with each having its own unique *despreading* sequence applied to separate it from the other signals. As each signal is separated with its despreading sequence, power in the other signals is spread out as if it were background noise. All spreading codes used must have the mutual (i.e., shared) property of *orthogonality*. That is, they must be noninterfering with each other. The practical limitation on the number of signals that can be transmitted in the same bandwidth is the number of orthogonal spreading codes available. In turn, the number of spreading codes depends on the number of bits in the spreading sequence: The longer the sequence, the more orthogonal codes exist. Spreading code design is a complex subject.

The type of spread spectrum system illustrated here is called *direct sequence* spread spectrum. There exists another practical spread spectrum technique, known as *frequency hopping* spread spectrum. In this technique, the frequency on which the digital signal is transmitted is moved multiple times during the transmission of one bit. The receiver follows the transmitter's spectrum hopping. One can see intuitively that a narrow band of interference present at one or a few of the frequencies in the hopping sequence will have little effect on the received signal.

With any spread spectrum approach, the total spectrum occupied by the signal is much wider than what would be occupied by the signal without spreading. Spectrum efficiency must be improved by transmitting many signals in the same spectrum at the same time.

4.4 Measuring Digitally Modulated Signals

Of great interest to cable engineers is the need to measure digitally modulated signals. Some measurements may be made using the same instruments as used in analog-modulated signals, such as a spectrum analyzer. Other measurements require new instrumentation.

4.4.1 Measuring Signal Level

Unlike an analog signal, a digitally modulated signal has no distinct peak in signal level at carrier frequencies. Most of the time, the carrier is not transmitted, at least not at full amplitude. Because digital signals are almost always randomized before being modulated, the spectrum is flat over its passband. As stated earlier, the average signal level is measured, not the peak level as is the accepted practice in measuring analog video signals. A spectrum analyzer is often used for this measurement, but several corrections in its reading must be made. The biggest is to correct for the occupied bandwidth of the signal, versus the bandwidth over which the measurement is made. The bandwidth over which the signal level is measured (the *noise bandwidth*) is roughly equal to the resolution bandwidth of the spectrum analyzer. However, the two are not quite the same, and for accurate measurements a correction factor must be used.

Since signal energy is distributed evenly across the band, the power of the signal is proportional to the bandwidth over which the measurement is made. This can be corrected to the actual bandwidth of the signal as follows:

$$S = 10 \log \frac{\mathrm{BW_{noise}}}{\mathrm{BW_{signal}}}$$

where

$\mathrm{BW_{noise}}$ = Noise bandwidth of analyzer IF filter
$\mathrm{BW_{signal}}$ = 3-dB bandwidth of signal

This corrects for the occupied bandwidth versus the measuring bandwidth, assuming that the spectrum is flat. Additional corrections that must be made include one for how the detector in the instrument responds to noise and one for how averaging is done. The instrument's manufacturer must specify these corrections. In most modern measurements the analyzer has a noise marker measurement facility that makes these corrections for you.

The noise bandwidth we refer to in the equations is defined as the bandwidth that would be occupied by an ideally flat filter with infinitely sharp edges (*skirts*), which admits the same amount of flat noise as does the real filter.

Note that the same measurement technique may be used with QAM, OFDM, or 8-VSB signals. With the 8-VSB signal, a small carrier component is transmitted 310 kHz above the band edge, but it is so small that it makes little difference in the signal level. Instruments other than a spectrum analyzer may be used to measure the amplitude of a digital signal, but the manufacturer must provide conversion factors to correct the measured reading to the actual average signal level.

4.4.2 Measuring Bit Error Rate (BER) and Modulation Error Ratio (MER)

The quality of a digital signal may be evaluated using *bit error rate* (BER) measurements if the signal quality is low enough to introduce bit errors. If the signal quality is such that errors are not being made, *modulation error ratio* (MER) may be used as an indicator of signal quality.

Bit Error Rate

BER is simply the average number of bits received in error divided by the total number of bits received. Scientific notation is used to express BER, since the number of errors had better be quite small. If one bit out of every million received is in error, then the BER is

$$\text{BER} = \frac{1 \text{ errored bit}}{10^6 \text{ bits}} = 10^{-6}$$

Note that BER is dimensionless. In order to measure BER in a laboratory setting, usually a known bit pattern is transmitted repeatedly. The receiver compares the received bit pattern against what it knows was transmitted and then displays the BER. It can take a long time to measure BER when the BER is low. Suppose it is desired to measure the BER of a 64-QAM signal being transmitted at a rate of 30.342 Mb/s. If the BER is 10^{-10}, then we will wait, on the average, through 10^{10} bits before we receive a bad one. This will require

$$\frac{10^{10} \text{ bits}}{30.342 \times 10^6 \text{ bits/sec}} = 329.6 \text{ sec} = 5.49 \text{ min}$$

before one error is received. We must receive a number of errors before we can conclude that we have a statistically valid sample. You can see that it will take a long time to make this measurement.

In-service BER cannot be performed on a known bit pattern because, by definition, the receiver cannot know what the transmitter is sending, unless the transmitter and receiver are located at the same place and the receiver is also getting the bit stream being sent to the transmitter. There are ways of estimating BER from the received signal without having to know the transmitted pattern. One way is to look at the eye diagram (see Figure 4.11). As more interference, noise, or other impairments enter the transmission path, the open center of the eye closes. The BER may be estimated from the amount of eye closure.

Another way to estimate BER is to look at the number of times errors are corrected. Most digital systems transmit extra bits for the purpose of correcting errors in the transmission (see Section 3.6.2). When these corrections are made, an error is counted. The bits received are also counted, and the ratio becomes the BER.

BER may be measured before or after error correction. These measurements are sometimes called the *raw*, or *precorrected, BER* (measured before error correction) and the *corrected*, or *postcorrection, BER* (after error correction). Both measurements have their place. Precorrected BER is the better indication of the channel performance, but postcorrection BER is the better indication of the quality of the signal the user will experience.

Modulation Error Ratio

Figure 4.22(a) plots bit error rate against transmission quality. The transmission quality includes degradation due to noise, distortion, frequency response errors, group delay, echo, gain compression, and other things that can damage the quality of the digital signal. In the left region, where transmission quality is low, the BER specifies the quality of the signal. However, above a certain noise density, the BER is so high we either don't see errors or at least don't have the patience to wait long enough to measure them. Even in this region, however, we are interested in knowing the quality of the channel. Since the change in transmission quality between a perfect picture and no picture is very small, it is possible that a technician will leave a home with perfect picture quality, only to have the subscriber call a few minutes later with no picture at all. Thus, it is not enough to know that a picture is good or that data is being transmitted efficiently. We need to know the quality of the channel even when things are going well.

A useful metric for communications quality is the *modulation error ratio* (MER). MER is defined in Figure 4.22(b), which shows one quadrant of the constellation diagram for a 16-QAM signal. Specialized test equipment is used to display the constellation diagram and to measure the MER. In Figure 4.22(b) the nominal location of each point in the constellation is shown as an *open circle*. Several pluses mark locations that the constellation point may occupy

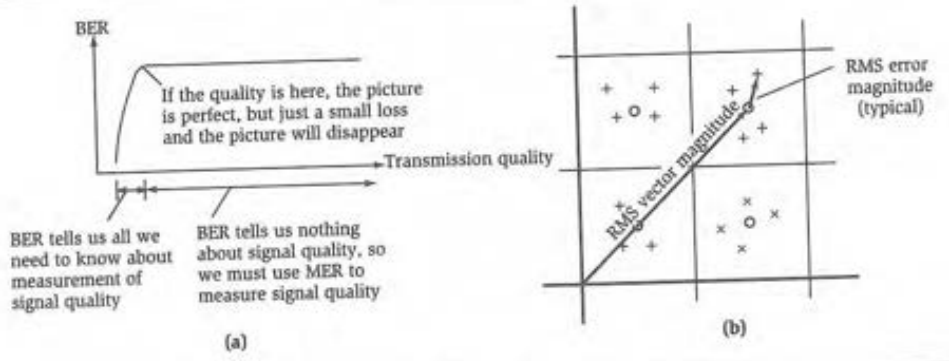


Figure 4.22 (a) BER vs. transmission quality. (b) Definition of MER.

for some one particular transmitted symbol. In practice, the whole area around the nominal location of the point will be occupied with points, because channel impairments affect each symbol differently. Some common channel impairments are shown in Figures 4.12 and 4.13. MER is a single number used to evaluate the channel quality regardless of the source of impairment.

MER is defined as 10 times the log of the ratio of the RMS error magnitude to the RMS vector magnitude to the nominal position of the point:

$$\text{MER} = 10 \log \frac{\text{RMS error magnitude}}{\text{RMS vector magnitude}}$$

The error magnitude and the vector magnitude are defined in Figure 4.22(b). Note that, while the RMS vector magnitude is constant for each point in the constellation, the RMS error magnitude will change for each symbol, since random processes are at work. In order to derive the MER, the test instrument sums the RMS error magnitude over many samples, at each point in the constellation, and averages them.

MER measures the "fuzziness" of the constellation. MER is sometimes described as similar to C/N measurements in analog transmission. While there are similarities (both are measured in decibels and higher numbers are better), there are also differences. Many things other than just poor C/N contribute to degradation of MER. Any transmission channel impairment affects MER, including random noise, group delay, echoes, phase noise, and gain compression.

In order to achieve good performance with 64-QAM, a MER of at least 23 dB is required. For 256-QAM, a MER of at least 28 dB is necessary. A minimum safety margin of 3 dB above these numbers is recommended. Portable instruments available tend to have a maximum BER measurement capability of about 34–38 dB.[9]

A related measurement you may see sometimes is the *error vector magnitude*, defined as the ratio of RMS error magnitude to the RMS vector magnitude, expressed as a percentage:

$$\text{EVM} = 100 \frac{\text{RMS error magnitude}}{\text{RMS vector magnitude}} \%$$

The impairments mentioned earlier have different effects on MER and BER. Noise (see Figure 4.17) is the most studied and is well documented. Phase noise and compression affect mostly the outer points of the constellation, so the MER is not affected much, yet BER may be degraded seriously. CW interference affects both MER and BER because all constellation points are impaired. Impulse noise creates havoc with the BER, but because of its transient nature the average impairment is small and MER is still good. Composite second order (CSO) and composite triple beat (CTB), because of their peaks, look more like impulse noise than CW interference. Each impairment has its own "power," and the MER reflects the sum of these powers.[10]

Because digital transmission is not perfect, transmission errors do occur. To render the system operational, *channel coding* is used to mitigate such transmission errors. A good example is the downstream digital video or cable modem transmission standard, ITU-T J-83 Annex B, which shows three levels of channel coding.

The best-known channel coding is *forward error correction* (FEC). The data words are grouped in packets, to which are appended Reed–Solomon (R-S) error-correcting words, e.g., 122 data words for six R-S words (7-bit words). At the receiver the R-S codes are used to detect almost all errors and, in addition, can correct one error for each two R-S codes (in this case up to three errors per packet). This is the basis of the pre-FEC and post-FEC bit error rates.

The second channel-coding mechanism is *interleaving*. Instead of transmitting all the words of each packet in sequence, they are interleaved. That is, adjacent transmitted words comes from different packets. For instance, all the first words of 128 packets, then all the second words, and so on until all the 128th words of all 128 packets are transmitted. At the receiver, the transmitted words are *deinterleaved*, to be regrouped in their original packet order. If impulse noise causes errors on many consecutive transmitted words, then without interleaving the R-S could not correct all errors. With interleaving, the grouped errors in transmission are now spread over a larger number of packets, each having only a few correctable errors. Interleaving provides a *burst protection period* proportional to the length and number of packets interleaved. On the other hand, it introduces a delay, or *latency*, in transmission and reception due to the storage requirement of the packets. This is why cable modem specifications limit interleave to 128×1. Digital video is more tolerant of latency, and deep interleave of 128×4 is often used (that is, the interleave is done over 128×4 packets, each of 128 words). The protection period is longer, at the expense of larger latency.

The third channel coding present in Annex B is *trellis code modulation* (TCM). TCM does not allow all combinations of symbol sequences. Rather, it limits the choice of the following symbol to specific ones according to the preceding symbols and specific rules. At reception, the same rules are used to select *the most probable word sequence*. There is no guarantee that it's the right one, but it is likely to be. Many errors are corrected, but some are not, activating the second line of defense, the Reed–Solomon FEC. Because it gives the most probable sequence but no decisive error indication, we cannot measure a pre-TCM bit error rate. Finally, Annex A and Annex C do not have TCM and consequently have a worst pre-FEC error rate than Annex B.[11]

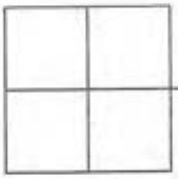
4.5 Summary

This chapter has presented the theory of digital modulation as used in cable TV and broadcast systems. The most commonly used family of modulation formats are quadrature amplitude modulation, QAM. The number preceding the QAM designation indicates the number of states the digitally modulated signal can

assume, and ranges from 4-QAM (also known as QPSK) to 256-QAM. There may be future applications for even higher levels of QAM. The next chapter will address a companion topic, protocols used in DOCSIS and other digital transmission standards of interest to cable operators.

Endnotes

1. Advanced Television Systems Committee, A/54, *Guide to the Use of the ATSC Digital Television Standard*, October 1995. Available at http://atsc.org/ (note: no "www").
2. Institute of Electrical and Electronics Engineers, *The Authoritative Dictionary of IEEE Standards Terms*, IEEE 100, 7th ed., 2000.
3. Society of Cable Telecommunications Engineers, ANSI/SCTE 22-1 2002, *Data-Over-Cable Service Interface Specification, DOCSIS 1.0 Radio Frequency Interface (RFI)*, Table 4-12.
4. R. D. Katznelson, *Delivering on the 256-QAM promises*. Cable-Tek Expo 2002.
5. Yasuhiro Ito, Performance of OFDM Transmission System — Peak-to-Average Power Ratio and Clipping Effect. *SPECS Technology*, June/July 1995, Cable Television Laboratories.
6. National Cable Television Association, *NCTA Recommended Practices: Upstream Transport Issues*. Though the citation is oriented to return path issues, the definition is generally applicable.
7. John G. Proakis, *Digital Communications*. New York: Mc Graw-Hill, 1983, p. 187.
8. Portions of this information are supplied by M. Bugajski of Arris Interactive, Duluth, GA., in a summary prepared for internal use.
9. Sunrise Telecom, *Modulation Error Ratio Demystified*, available at www.sunrisetelecom.com.
10. Sunrise Telecom Broadband On-line Learning, *MER, BER and More*, available at www.sunrisetelecom.com.
11. ITU-T J-83, Digital Multi-Programme Systems for the Television and Data Services for Cable Distribution.



Chapter 5
Cable Networking Protocols

5.1 Introduction to Protocols

This chapter introduces you to the concept of data protocols, with an emphasis on those used in DOCSIS modems. A number of other protocols are covered because they find frequent use in cable TV work, but they were not developed specifically for that. You will not become a protocol expert by reading this chapter, but you should be able to understand the basic concepts.

The word *protocol* is generally not used in analog communications, but it could be applied. When two pieces of equipment communicate with each other, there must be an established set of rules for that communication. In digital communications, the set of rules is called a *protocol*. For example, if we were to apply the word to the analog world, we might say that the protocol for FM radio broadcasting in monaural would be as follows:

Deviation: 75 kHz peak

Preemphasis: 75 μs

Frequency: channels spaced 200 kHz apart, beginning at 88.1 MHz, through 107.9 MHz (North America)

This would define reasonably well a protocol to which monaural FM receivers could be designed, with assurance that a receiver built by one company would be compatible with a transmitter built by someone else. Data communications protocols are more complex but are intended for the same purpose: to make sure that a talker will be understood by the listener.

5.1.1 The Concept of Protocols

Data protocols can be introduced using the analogy of a letter. A letter is written by one person and intended to be read by another. In the process of sending the letter from the writer to the reader, it is handled by a number of different people and packaged in different ways with other letters. Those carrying the letter have no interest in, and no need to know, the content of the letter, which is "hidden"

in an envelope. Similarly, a digital message from one entity to another is "packaged" with other messages and transported by a transport system that has no knowledge of the contents of the message.

As shown in Figure 5.1, the sender of a letter addresses it to the intended recipient. The letter is placed in a bag with other letters for transport to a sorting center and then rebagged with other letters going to the same general destination. The bag is placed on a conveyance (truck, train, or airplane), which carries it close to its destination, where the process is reversed, before the letter is ultimately delivered to the intended recipient, who removes it from the envelope and reads it. In data transmission, the "envelope" is replaced, typically, with a *header* and sometimes a *postheader*. Subsequent handling steps pay no attention to what is enclosed, nor would they know what to do with the contents if they did look.

Just as a letter can be placed in an envelope, the envelope placed in a bag, and the bag placed in a train, different layers of the protocol may place their own header and postheader on the message. Those headers are interpreted by the corresponding process at the receiving point and discarded. The message is then passed to the next process.

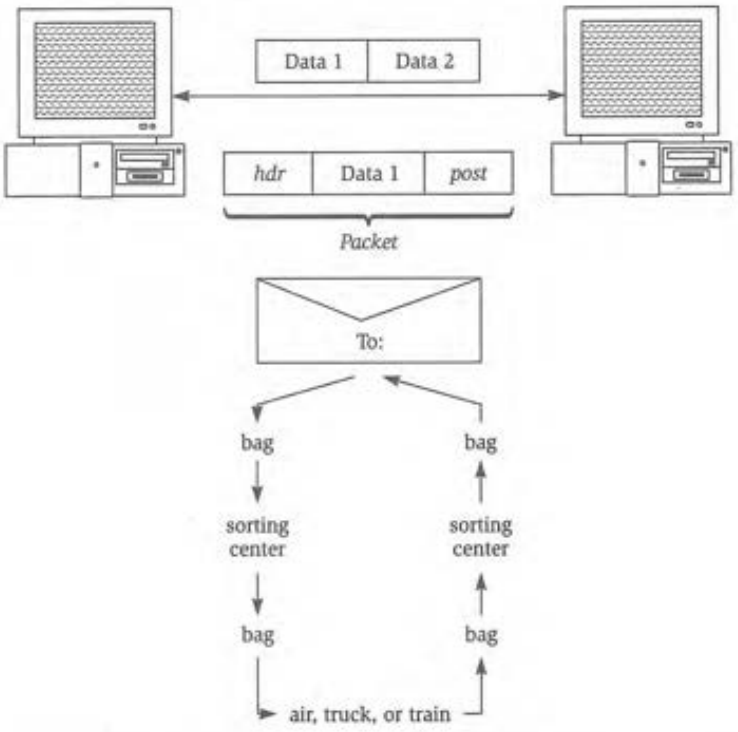


Figure 5.1 Concept of a data protocol.

A communications protocol defines the set of steps a message takes from the sender to the receiver. Analogous to the letter concept of Figure 5.1, the protocol defines how a data signal is handled at each stage of transmission. One layer knows nothing about how the message is transmitted at a different layer. Just as a letter-sorting process has no interest in whether the letter is to be transported by airplane or truck, one layer of a protocol has no interest in how the message is to be handled at another level.

One advantage of using the layering concept is that large portions of the system can be designed such that data can be carried on a number of different media. For instance, a data message can be constructed the same whether it is to be sent over a short point-to-point wired link or over a satellite link shared with a lot of other users. This enables efficiencies in writing software and constructing hardware.

5.1.2 The OSI Seven-Layer Reference Model

A common way to explain the role of different protocol levels is the so-called seven layer model promulgated by the International Telecommunications Union (ITU). The model is known as the *O*pen *S*ystems *I*nterconnection (OSI) model.[1-3] 'ISO' is sometimes used, for the International Standards Organization. Not all modern protocols fit neatly into the seven layers. Nonetheless, the model provides a useful frame of reference for understanding protocols.

Figure 5.2 illustrates the OSI seven-layer model and includes examples of common protocols and approximately where they would fit into the model. The model involves transmitting information from, for example, the left column (sender) to the right column (recipient). At the top of each column ("stack") is the application, such as a Web browser or an email system. The bottom three or four layers are usually of the greatest interest to engineers working on data transmission. These bottom layers define the transmission network, whereas the higher layers define the end user system.

You can think of Figure 5.2 as describing the flow of information in a computer or other device, from the application (Web browsing, electronic program guide, etc.) through to the communications port on the device. The information is transported to whatever machine is talking to the first and flows up again. In the middle is the communications network, which involves up to layer 3 or maybe 4. Note that different protocols may exist in the end applications and the communications system. The same protocol is not necessarily used in the end machines as in the communications network. This is fine as long as the devices can interface using a common protocol. Later we describe some of the more common transport protocols you will encounter.

Physical Layer

The lowest layer is the *physical* layer, which defines the functions required in order to get data from one place to another. Besides specifying such things as the

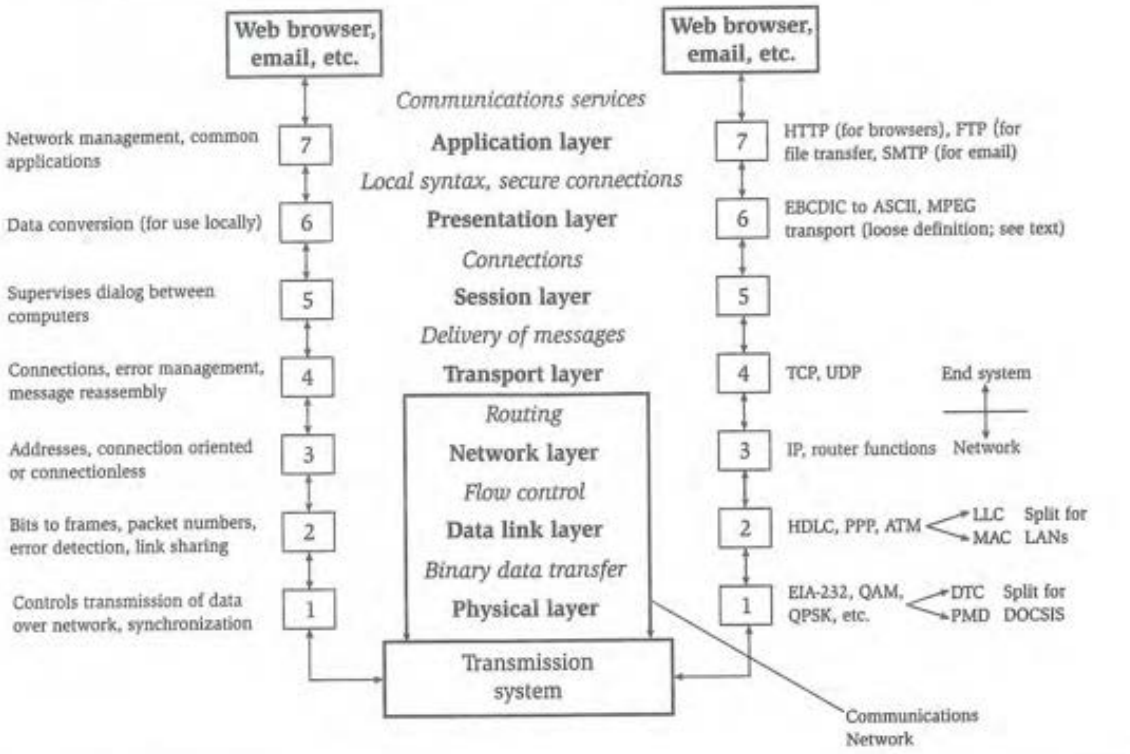


Figure 5.2 The OSI seven-layer model.

modulation format (to which Chapter 4 was dedicated), the physical layer takes care of such functions as synchronization. *Synchronization* refers to the process of ensuring that data recovery at the receiver is timed properly with the data being transmitted. The actual transmission system, such as an HFC network, is not considered part of any layer of the protocol, but obviously the protocol must have knowledge of the function below (and above) itself.

Data Link Layer

The *data link layer* handles such tasks as gathering up sets of bits for transmission as packets and making sure the packets get from one end to the other. In addition, recognizing that physical layer transmission sometimes introduces errors, the data link layer handles error detection (and sometimes correction).

When the OSI model was first introduced, data communication was point to point. With time, data communication has become much more multipoint to multipoint, so the data link layer was divided to recognize multipoint needs. With the split of the data link layer, the traditional functions of the layer went to the upper portion, called the *logical link layer*. A new sublayer, media access control (MAC), was defined. The MAC layer routes packets from a sender to a receiver along a common path. It makes sure the message arrives at the intended recipient. The MAC layer adds a physical address, defining the intended recipient machine, and controls shared access to a resource. For example, control of TDMA operation (see Section 4.3.1) is a MAC layer function.

Network Layer

The network layer handles routing of packets over a network that may consist of many links. It figures out, for each packet, the best of the available routes over which to send a packet. This is explained in more detail in Section 5.1.4.

Transport Layer

The transport layer traditionally is an end user layer and not a part of the communications network. However, as we shall see, modern communications networks are often cognizant of the layer 4 protocol, at least in order to allow for quality-of-service handling of packets.

End System Layers

The upper layers of the protocol are important, but they are treated lightly here because the emphasis is on the lower layers of the protocol, with which communications engineers tend to interact more frequently. The transport layer combines multiple low-data-rate channels into a higher-data-rate channel and then breaks down the channels at the receive end. It defines the destination for each channel. In a connection-oriented service, the transport layer sets up the connection, delivers error-free messages in the correct order, and ends the connection when it is no longer needed. In a connectionless service, the transport layer delivers messages one by one, possibly with errors and with no guarantee of the correct order.

The session layer supervises the dialog between two computers. It is used, for example, in the transmission of audio and video channels in a video conferencing session. Synchronization points are inserted between large numbers of messages, such as the continuous messages of video and audio files. Synchronization points are packets that divide the sequence of messages into groups.

The presentation layer converts between different representations of data. For example, some computers represent words by placing the most significant

bit first, whereas other computers do the opposite. The presentation layer adapts the data to be used on the host machine if necessary. MPEG video encoding may be loosely thought of as a presentation layer function. The MPEG transport layer probably would be better thought of as a combination of the transport layer and some functions from lower layers. (MPEG does not really conform to the seven-layer model.)

The application layer interfaces with end application programs that require communications functions. The application programs, such as Web browsers, don't *reside* in the applications layer themselves; they are above all the communications services. The application layer provides file transfer, terminal emulation, remote log-in, directory service, and remote job execution. These services are used by the end application.

5.1.3 MAC Layer Issues of Interest in Cable Modem Applications

Of particular interest to users of cable modems is the MAC sublayer of the data link layer. It falls to the MAC layer to provide an efficient protocol to allow operation of modems in the HFC environment. This physical environment is different from that of many traditional data networks. A significant issue is that each upstream "talker" cannot "hear" other upstream talkers unless the upstream transmission is repeated in the downstream direction. Also, unlike local area networks, the propagation delay on the HFC plant is significant and, if not taken into account, can create inefficiencies in data communications. The following discussion of the use of time slots presupposes a time division multiple access (TDMA) protocol in the upstream direction. It serves as an introduction to the more detailed section 5.2 on cable modems, which follows.

Reservation of Time Slots

In TDMA systems, some method must exist to allow for determination of which of many transmitters gets to transmit in a given time slot. One such method is generally known as *reservation*. In a reservation protocol, each talker who wants to talk requests, and is then assigned, one of the available time slots. When that time slot arrives, he or she talks and then is quiet while others talk. The management and assignment of the time slots comes from some central authority, which in the case of a cable modem system would be the headend *cable modem termination system*, CMTS. A reserved time slot allows a modem to transmit a known amount of information in a known time period. However, if a modem does not wish to use its time slot, the time slot is not used by anyone else until the controller decides that the time slot is available and assigns it to someone else. This can result in inefficiencies when there are a lot of modems, each with little to say. On the other hand, when a modem has a lot to say, reservation can be more efficient than the alternatives.

Contention for Time Slots

An alternative is to have the modems contend for time slots: When a modem needs a time slot, it simply speaks up and requests the time slot. In the case of cable modems, where one modem cannot "hear" any other modems, the talking modem will have to wait until the headend can acknowledge that it received the transmission. If another modem tried to talk at the same time — even if only part of the transmissions overlapped — then neither would be heard, so neither would be acknowledged. The transmissions are said to have *collided*. When that happens, both modems wait for an arbitrary time and then try to transmit again. If they again collide or if one collides with any other modems, they must again back off for a random time and try yet again. With some protocols, such as Ethernet, the talkers can hear each other, so they know when a collision occurs.

When talkers have relatively little to say but need to say it quickly when they do have something, contention works fairly well. In cable systems, though, contention is particularly difficult, because one talker cannot hear other talkers unless the upstream transmission is repeated in the downstream direction. This was done in some early (pre-DOCSIS) cable modem systems, which were adaptations of Ethernet. However, the propagation delay of a round-trip on cable plant can be long compared with the length of a message. Consequently, by the time two modems know that they have collided, a lot of time has been wasted. In addition, the need for upstream data to be repeated in the downstream direction is an inefficient use of the spectrum resource.

Combination Protocols

Contention protocols are sometimes called *aloha* protocols. Simple aloha protocols do not feature time slots, but they do allow any modems that need to talk to do so at any time. If the protocol follows the discipline of time slots into which any talker's transmission falls, it is called a *slotted aloha* protocol. If some time slots are available for reservation while others are available for contention, the protocol might be called a *reserved aloha* protocol.

The difficulties with contention systems on cable might lead one to prefer some sort of reservation system. It is certainly possible to reserve for each modem on a particular channel a permanent time slot for upstream transmission. However, such a permanent assignment is a problem when each modem is in use for at most a few hours each day and has little to say when it is used. The rest of the time, the time slot cannot be occupied by other users. Some sort of method should be used to assign time slots on an as-needed basis so that if a modem was not using its time slot, that slot could be used by another modem.

If a modem is assigned a time slot only when it needs one, then there must be some way provided to allow a modem to request a time slot. Since the headend can have no a priori knowledge of when a particular modem will need a time slot, it is necessary to allow a modem to request one. This request implies

that there should be some mechanism of contention in a cable modem protocol. The mechanism can be one where a time slot is allocated to all modems not presently talking, in which they can request a time slot. They must do this in a contention mode. A modem needing a time slot waits until the contention time slot, and then makes a request to the headend. If it is the only modem doing so, its request will be acknowledged in the return datastream, and a time slot will be assigned if available. If a second modem has also transmitted a request at the same time, then neither will be acknowledged, and they must both back off, waiting for a random time before trying again in the contention time slot.

5.1.4 Connection-Oriented and Connectionless Service

A communication link may be differentiated on the basis of whether it is a connection-oriented service or a connectionless service. Connection-oriented services are better known to the public, as represented by the public switched telephone network (PSTN). Connectionless service is a less familiar concept. Both are described here.

Connection-Oriented Service

Figure 5.3 represents a connection-oriented service, such as telephone service. It shows a section of a switch, which couples one user (users A through H are shown) with any other user. A switch may be thought of as a matrix, with users connected to horizontal lines and vertical lines used to interconnect users as

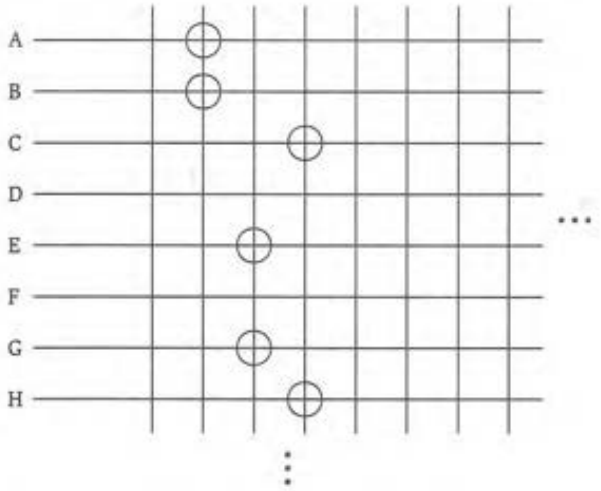


Figure 5.3 Representation of a connection-oriented service.

demanded. The circles represent connections between the horizontal and vertical lines. As shown, A is connected to B, C to H, and E to G. D and F are not in communication with any users.

A connection must be set up whenever communication is desired between any two points and then "torn down" when the communication is completed. Because time is required to set up (and tear down) a communications channel, the user must wait until a channel is set up before he can use it. Resources are dedicated to the communications channel whether or not communication is taking place at the moment. For example, during a voice telephone call, a channel is dedicated to that call even when one party has left the phone to get some information. Resources are tied up for both directions of the call even though each direction is used no more than half the time.

Connectionless Service

By contrast, a connectionless system passes information when information is available; the rest of the time, the resources are available for other users. Each packet, or "chunk", of information is treated on its own, without relation to anything else. There is no setup or teardown of a connection. Figure 5.4 illustrates the concept of routing, a key component of a connectionless circuit. The

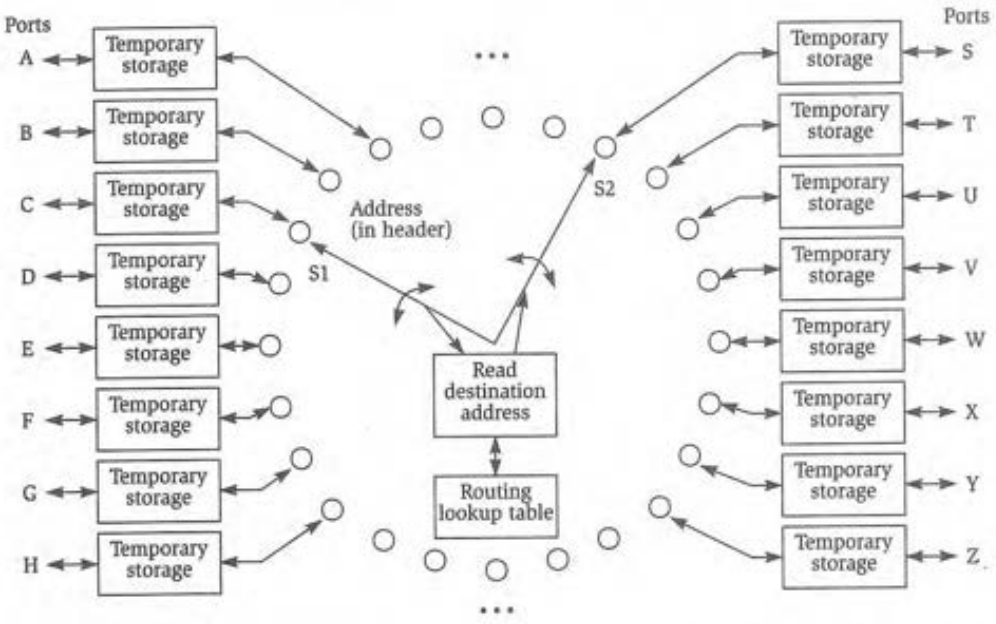


Figure 5.4 Router, representing a connectionless system.

concept is used in the Internet backbone and other IP systems. The figure shows communication ports A through H and S through Z, each connected to a temporary storage facility, which accepts a message from the user and stores it for a short time until it can be routed to the next destination. Each interface can be thought of as either a user or another point in a multipoint route between two endpoints (users). All communications routes are two way.

A rotary switch has two commutators, S1 and S2, each of which can independently connect to any interface. Conceptually, S1 connects in turn to each port. If there is data there to be routed further, that data will contain a header that defines the destination. The destination header is read, and a routing lookup table is consulted. The lookup table tells the router which of the ports moves the signal most expeditiously toward its destination. Commutator S2 connects to this port, moving the signal toward the next point on the journey to its destination. (Real routers don't use switches to make connections, but this is a convenient mechanism with which to illustrate the function.)

If A is talking to B (as in Figure 5.4), then at some time S1 will connect to A, the destination address B will be read, and the routing lookup table will be consulted. In this case, the lookup table would say to connect to B directly, since the destination port is connected to the router.

Figure 5.5 shows a larger view of a network. Each of the circled letters represents a router, as in Figure 5.4, each in a different location. Each router is connected to a number of adjacent routers. The routing table in each must know of every ultimate destination and must know how to efficiently route a packet toward that destination (this concept is modified in the case of some modern protocols, such as MPLS). In several cases, multiple routes are possible, so if one

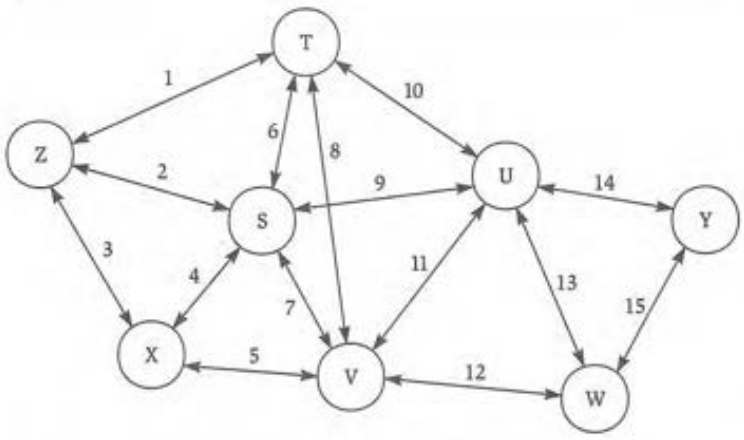


Figure 5.5 Multiple point routing.

is not available another can be used. For example, to send a packet from Z to X, the most efficient route is 3, which connects the two directly. If route 3 is not available, then the packet could be sent to S via route 2 and then to X via route 4. Route 3 might not be available because it has failed, been taken out of service, or lacks the capacity to handle all of the traffic being fed to it at the moment. The routing table in each router must be aware of the best route and alternates, and must be aware of the present condition of each route.[4]

Differences Between Connection-Oriented and Connectionless Service

In summary, connection-oriented services, an example being the public switched telephone network, dedicate a communications path to two users for the duration of a communications session. The session may last less than a second, or it may last for days. When communication is desired, a circuit must be set up that is dedicated to that communicating pair for the duration. At the end of the session, the circuit is torn down, releasing any common resources for other users. The device at the heart of a connection–oriented service is called a *circuit switch*.

In contrast, connectionless circuits are established on an ad hoc basis when they are needed for an individual packet of information. As soon as the packet passes, the resources are available for other users. Each packet of information has a destination address. That destination address is read by the router, which then consults a routing table to learn the best route on which to send the packet at that moment.

To the user of a connection-oriented service, it is necessary to establish the circuit (e.g., by placing a phone call), wait for the connection to be established, and then use it. The delay, once the circuit is established, is short and predictable. A user of a connectionless service is not aware of the establishment of a path for each packet, that function being handled automatically in protocol. So far as the user is concerned, the communication path is always available to him. On the other hand, because of route diversity, sequential packets may travel on different paths; and since some routes may be faster than others, it is not guaranteed that packets will arrive on a regular basis or even in the right order. It is the responsibility of the protocol to rearrange packets into the correct order and to present them to the using function correctly. (With reference to the OSI model of Figure 5.2, this function would be handled by the transport layer.)

Connectionless services use system resources more efficiently because no resources are tied up between users who have established a communications path they are not actively using at the moment. On the other hand, the user of a connectionless service "thinks" a communications channel is available at all times, though he or she may experience delays. Of course, many of the delays are small in human terms, and packets received out of order may be properly ordered without user awareness.

5.2 The DOCSIS Protocol for Cable Modems

Early cable modem systems used proprietary protocols, many derived from Ethernet. A group of cable television operators and Cable Television Laboratories (CableLabs) collaborated in the development of a set of purchase specifications to be used by participants and other interested parties as a procurement standard for cable modems. The collaboration initially operated under the name MCNS Holdings, L.P., an acronym for Multimedia Cable Network System. The set of specifications is referred to as DOCSIS, for the formal name of the specification set, data-over-cable service interface specification.[5] The DOCSIS standards were later adopted by the Society of Cable Telecommunications Engineers (SCTE), a recognized standards body. Subsequently they were adopted as national standards by the American National Standards Institute (ANSI).

DOCSIS was the first opportunity to allow cable TV equipment to be owned by the consumer. Standardization is necessary to ensure that any modem bought by any consumer can connect to the network. Further, the subscriber can assume that if he or she moves to another cable system, the same modem will work in the new system. The standard allows multiple vendors to make competing products. With competition comes innovation and lower costs.

5.2.1 DOCSIS Reference Model

Figure 5.6 illustrates the headend-to-home portion of the DOCSIS reference architecture.[6] The architecture is fairly straightforward. In the downstream

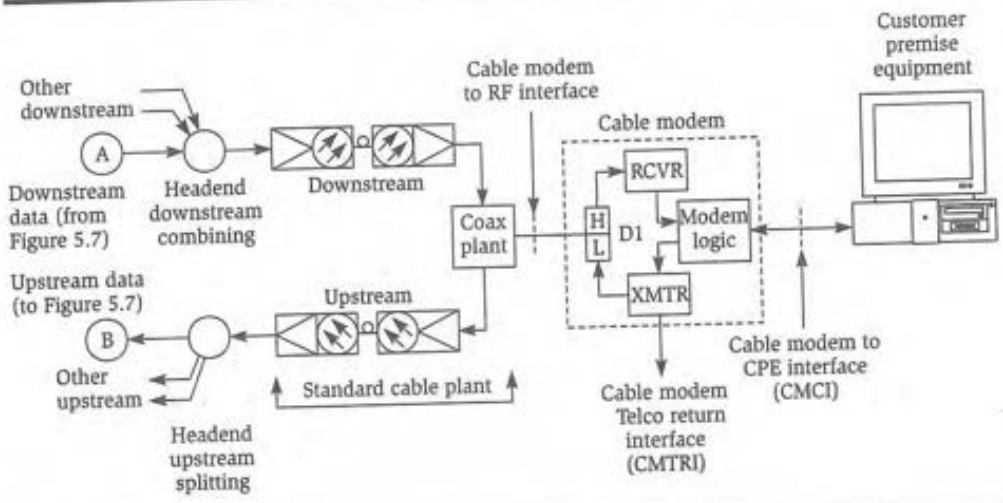


Figure 5.6 DOCSIS reference architecture, headend to home.

direction, the data signals, bubble A, are coupled with other downstream signals and supplied to the downstream HFC network (Chapter 9 of this book covers headend RF architectures in more detail). At the cable modem, the signal is supplied to diplex filter D1, which separates the downstream signal (above 54 MHz) from the upstream signal (below 42 MHz).

The downstream signal is supplied to the receiver in the cable modem, whose output is the downstream data. In turn, that data is supplied to interface logic, which converts the signal to one of several defined interface protocols. One such interface is the 10baseT protocol, described later in this chapter. This protocol is a particular physical layer implementation of Ethernet, which operates at a data rate of 10 Mb/s using four conductors from a CAT-5 cable and RJ-45 connectors (the interconnecting cable looks like a telephone cable with bigger connectors). It has been selected as a standard interface because of its suitable speed and the availability of 10baseT interfaces in many computers. For computers lacking built-in 10baseT interfaces, suitable add-in cards are available.

Two other interfaces are defined for DOCSIS-compliant modems. One is the universal serial bus (USB).[7] The other is the peripheral component interface (PCI).[8] USB has gained acceptance as an interface between computers and a number of external devices. The original USB specification is a 12-Mb/s bus that allows many external devices to be connected to the PC via a hub. The more recent USB specification operates at speed of 480 Mb/s. The PCI interface is used with internal modems and other hardware. The original specification has a maximum transfer rate of 132 MB/s, and is backwards compatible with the older ISA bus. A more recent version of the specification enables transfer rates up to 4.26 Gb/s.[9]

A cable modem transmitter handles return signals bound for the headend. The cable modem acts as an Ethernet bridge at the MAC layer (see later for definition of the MAC layer). This means that it passes traffic from and to the user's computer but does not pass the remainder of the traffic on the cable. Thus, a user doesn't see traffic bound to other users on the node.

An optional telephone return interface is shown in Figure 5.6. Some early systems used telephone return, with data cable-delivered to the subscriber. This allowed modem service on cable systems that had not upgraded to two-way capability.

Figure 5.7 illustrates the headend portion of the reference architecture. Bubbles A and B correspond to the same bubbles in the previous figure. The *cable modem termination system* (CMTS) includes the demodulator and modulator, which interface to the cable system. It also includes the interface to the network. Several options are defined for interfacing with the Internet backbone, either locally or (probably more commonly) by reaching a remote Internet backbone point of presence via a wide area network. The cable modem termination system also interfaces with a security controller and an *operational support system* (OSS).

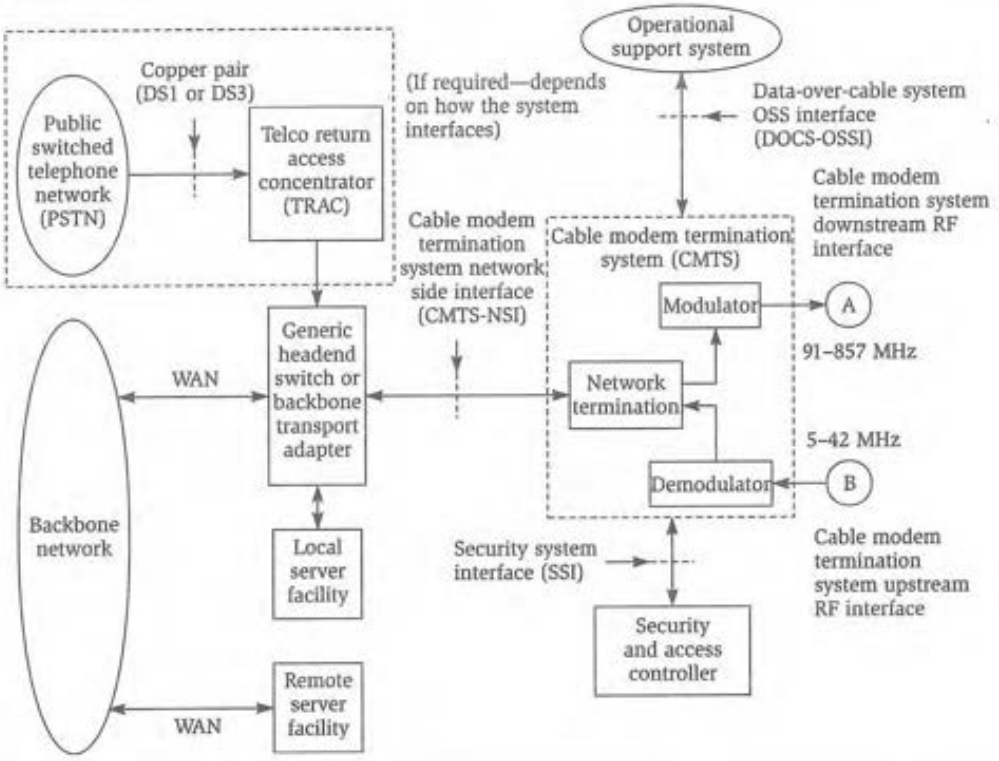


Figure 5.7 DOCSIS reference architecture, headend.

5.2.2 DOCSIS RF Characteristics

Table 5.1 illustrates the RF characteristics of the three versions of the DOCSIS specification. Downstream modulation is the same in all three specifications, being either 64- or 256-QAM, with the characteristics shown. Note that the symbol rate in the 256-QAM mode is higher than it is in the 64-QAM mode but that the excess bandwidth is lower (12% vs. 18%), resulting in a similar occupied bandwidth.

The upstream specification is identical for both DOCSIS 1.0 and 1.1, but in DOCSIS 1.0 the 16-QAM mode was not required and, in fact, was not implemented very often. Thus, the maximum speed of which a DOCSIS 1.0 modem was capable was twice the symbol rate of 2,560 kbaud (thousand symbols per second), equal to 5.12 Mb/s (two bits per symbol in QPSK). This is wire speed, which includes all overhead. In DOCSIS 1.1 the 16-QAM mode is required, so the maximum upstream data rate is twice what it is in DOCSIS 1.0, even though the specifications read identically on this point.

Table 5.1 Comparison of Cable Modem RF Features

Parameter	*DOCSIS 1.0*	*DOCSIS 1.1*	*DOCSIS 2.0*
Downstream			
Frequency (MHz)	91–857 MHz		
Modulation	64-QAM, 256-QAM		
Filter (each end)	$\alpha = 18\%$ root raised cosine (64-QAM) $\alpha = 12\%$ root raised cosine (256-QAM)		
Symbol rates (Mbaud)	5.056941 (64-QAM) 5.360537 (256-QAM)		
Corresponding bit rate (Mb/s)	30.342 (64-QAM) 42.884 (256-QAM)		
Upstream			
Frequency	5–42 MHz (65 MHz maximum, Euro DOCSIS)		
Modulation	QPSK (16-QAM optional, rarely implemented)	QPSK, 16-QAM	Same as DOCSIS 1.1 plus 8-QAM,* 32-QAM,* 64-QAM (S-CDMA and TDMA), 128-QAM (S-CDMA only)
Filter	$\alpha = 25\%$ root raised cosine		
TDMA mode symbol rates (kbaud)	160, 320, 640, 1280, 2560		Same as DOCSIS 1.1 plus 5120
S-CDMA mode symbol rates (kbaud)	Not applicable		1280, 2560, 5120
Channel widths (kHz) @−30 dB	200, 400, 800, 1200, 3200		Same as DOCSIS 1.1 plus 6400
Data rates (kb/s)	320, 640, 1280, 2560, 5120 (QPSK)	Same as DOCSIS 1.0 plus 10,240 (16-QAM)	Same as DOCSIS 1.1 plus up to 30.72 Mb/s in advanced modes
Upstream Preequalization	None	8 taps (opt.)	24 taps

* May not be supported at CMTS.

Early studies of cable modem data indicated a relatively symmetrical data pattern. DOCSIS 2.0 allows a higher bandwidth upstream channel, to approach symmetrical data rates. Two advanced modulation methods were incorporated. One is an advanced TDMA (A-TDMA), with 64-QAM modulation (8-QAM and 32-QAM are optional at the CMTS but required at the modem). Except for the filter excess bandwidth and the use of TDMA, this 64-QAM upstream modulation is identical to 64-QAM downstream modulation. TDMA, you will recall, applies to multiple transmitters, each of which takes some of the upstream transmit time and takes the entire channel; no other modem may transmit at the same time.

The other advanced modulation method is S-CDMA, in which each modem is assigned a spreading code, a pseudorandom sequence used to spread the carrier (see Section 4.3.4). Each modem transmits continuously using its unique spreading code, but at a lower instantaneous speed. Different modems are discriminated by their spreading codes.

A feature introduced in DOCSIS 1.1 and expanded in DOCSIS 2.0 is the use of a preequalizer in the upstream direction. Since the upstream signal is subjected to possible variations in amplitude and group delay as well as microreflections, the CMTS measures impairments and adjusts a tapped filter in the modem to compensate for the impairments. The filter is initially adjusted when the modem is first activated, and is continually updated as the modem responds to the headend. Thus, the filter is termed an *adaptive filter*. Since each modem communicates to the headend through a different path length, and that path is subject to unique distortions, it is necessary to use a different adaptive filter for each modem.

DOCSIS is backward compatible. DOCSIS 1.0 modems will work with a DOCSIS 1.1 or a DOCSIS 2.0 CMTS. DOCSIS 1.1 modems will work with a DOCSIS 2.0 CMTS.

Additional Downstream Physical Layer Features

The downstream physical layer is divided into two sublayers (see Figure 5.2). The parameters described so far reside in the physical media–dependent (PMD) sublayer. On top of the PMD sublayer resides the downstream transmission convergence (DTC) sublayer, which conforms to MPEG-2 (ITU-T Recommendation H.222.0). The DTC sublayer allows for different types of data, such as digital video streams, to be delivered to modems. Virtually any data service desired can be packaged within MPEG packets. The DOCSIS documentation doesn't define these services.

The PMD sublayer, the lowest layer, specifies, in addition to the parameters shown earlier, concatenation of Reed–Soloman block code and Trellis code for error detection and correction. Variable-depth interleaving supports both latency-sensitive and -insensitive data. Section 3.6.2 describes error correction.

Additional Upstream Physical Layer Features

The physical layer in the upstream direction includes other PMD features. The access method is time division multiple access (TDMA — explained in Section 4.1.3), in which a number of modems use the same channel at different times. The standard supports both fixed- and variable-length messages, to improve the efficiency of the return path. Modems may be assigned a time slot, or they may contend for an open time slot. Programmable Reed–Soloman block coding allows trade-offs in error correction and overhead. The standard provides means for controlling the frequency of the modem, the time at which it transmits, and the power level with which it transmits. (Section 16.6.2 contains information related to upstream level control.)

5.2.3 Data Link Layer—MAC

The lowest sublayer of the data link layer is the media access control (MAC) sublayer. The MAC protocol includes the following services:

- Bandwidth allocation, controlled by the cable modem termination system (CMTS).
- A stream of minislots in the upstream direction
- Dynamic mix of contention- and reservation-based upstream transmit
- Extensions provided for future support of ATM (asynchronous transfer mode) or other data PDUs (packet data units)
- Class-of-service support
- Extensions for security as well as virtual LANs at the data link layer
- Support for a wide range of data rates

Bandwidth allocation refers to the process of determining the data rates and frequencies to use in communications. This is done by the modem headend system, the CMTS, based on programmed parameters and existing conditions.

In order to achieve improved efficiency, and acknowledging that many upstream messages are very short, the protocol permits upstream time slots to be subdivided, allowing access by a larger number of modems. Also, data may be sent in packets of variable length, rather than requiring the same length in all packets regardless of the length of the message contained.

5.2.4 Data Link Layer—LLC and Higher Layers

Above the link layer security sublayer, the DOCSIS standard merges with other standards and so is not specified completely in DOCSIS documentation. The logical link control sublayer (class 1 only) is provided in accordance with ISO/IEC 10039. The network layer is the internet protocol (IP), initially version 4,

with an anticipated migration to version 6 as it becomes deployed. Services residing above the network layer use IP as a bearer, and the services are transparent to the cable modem system.

In addition, there are several network management and operation capabilities specified that depend upon the network layer, including the following:

- SNMP (simple network management protocol) for network management
- TFTP (trivial file transfer protocol), a file transfer protocol for downloading software and configuration information
- DHCP (dynamic host configuration protocol), a framework for passing configuration information to hosts on a TCP/IP network
- A security management protocol as defined by DOCSIS

5.2.5 Quality of Service (QoS) and Security

Table 5.2 compares the quality of service (QoS) features and the security of the modem standards. DOCSIS 1.1 and 2.0 are identical in these respects, the difference being in upstream modulation and data rates. The major QoS difference between DOCSIS 1.0 and DOCSIS 1.1 is that DOCSIS 1.0 uses service ID (SID) to identify cable modems and the devices behind them, while DOCSIS 1.1 uses service flows. DOCSIS 1.1 also has improved MAC framing features and improved provisioning and authorization, with advanced BPI+ features.

Quality of Service

Regarding QoS, DOCSIS 1.0 is a best-effort service, meaning that packets are handled when they can be handled. The operator can limit the maximum downstream and upstream data rates and can guarantee a minimum upstream rate. Other than that, there is little the operator can do to control data transmission through a modem. There is no way to tell what the traffic is and no way to limit it or to treat different types of packets differently.

The situation is much different for DOCSIS 1.1/2.0. The concept of *service flows* is introduced as a key element to the operation of the MAC layer. Service flows provide a mechanism for upstream and downstream QoS management. They are integral to bandwidth allocation. A service flow (SF) is a unidirectional datastream for one or more applications, which are to be treated in a prescribed manner during transport. Service flows are described in more detail in Section 5.2.6.

Payload header suppression (PHS) is a technique introduced in DOCSIS 1.1 for improving transmission efficiency, particularly with small packets, such as those used in voice on Internet Protocol (VoIP). The header on each packet can be quite long—on the order of 40 bytes. Compared with a maximum-length

Table 5.2 Comparison of DOCSIS QoS and Security

Parameter	*DOCSIS 1.0*	*DOCSIS 1.1 and DOCSIS 2.0*
Quality of service	Single tier, best effort, 2-way QoS (class of service) profile	Multiple static and dynamic unidirectional service flows with service class names; service-level enforcement statistics
Payload header suppression	No	Yes
Upstream scheduling types	No	BE, UGS, rtPS, UGS/AD, nrtPS
Dynamic service	No	Dynamic service add (DSA), change (DSC), delete (DSD)
Classifiers	No	MAC addr, 802.1Q, P, Ether type, IP addr, IP type, Port #, TOS
Fragmentation	No	Yes
Security	Baseline privacy (BPI)	Baseline privacy plus
Authentication	MAC address checked	X.509 certificate in modem
Key exchange	RSA algorithm with ECB mode of DES for key encryption; keys have a limited lifetime, must be renewed before expiration	X.509 certificates with triple DES; authorization key generated in CMTS, from which traffic encryption keys are derived; sent to modem encrypted with public key; keys have a limited lifetime, must be renewed before expiration
PDU encryption	CBC mode of DES for encrypting PDUs in both directions; keys have limited lifetime	Same as DOCSIS 1.0; multiple security associations for different data flows

Ethernet packet of 1500 bytes, this is not much. But many if not most applications will not be using 1500-byte packets. A relatively slow application would be delayed unacceptably if the transmission system waited until 1500 bytes of data were available. Thus, most packets are much shorter, as short as 64 bytes. In such a small packet, 40 bytes of header almost doubles the length of what must be sent. In order to lessen the overhead burden, DOCSIS 1.1 and later defines payload header suppression.

The header is identical for all packets related to a particular service flow, so for that service flow, the sender suppresses the header, substituting a payload header suppression index, an 8-bit field that identifies the suppressed header (previously agreed between the sender and the receiver). When the packet

reaches the receiver, it looks at the payload suppression index to determine the header that should have been sent and then adds it back to the packet.

Dynamic service will be illustrated in Section 5.2.6. The term refers to the ability of a DOCSIS 1.1 or 2.0 system to provide different services to different classifications of packets. These services may be set up on demand, within the subscriber's authorization. Dynamic services are service *flows* which are created, modified, or deleted after the cable modem registers. They do not force the cable modem to reboot. Compare this to *static service flows* which are defined by a cable modem configuration file installed during registration and after a cable modem reboots. Dynamic service actions include DSA (add), DSC (change), and DSD (delete), which are generically known as DSx.

Upstream scheduling types are also unique to DOCSIS 1.1 or 2.0. These special scheduling types augment the best-effort (BE) scheduling type in DOCSIS 1.0 and are applied to individual upstream service flows by the cable modem under control of the CMTS.

Unsolicited grant services (UGS) allow a modem to transmit or receive packets of data of a fixed size at a regular interval without making a request for each transmission. This is important, for example, in VoIP service, where you cannot tolerate much jitter in the arrival time of packets and the amount of data transmitted at each interval is the same. *Real-time polling services* (rtPS) allow polling of a single modem to see if it has data to send, permitting much more rapid response to a need to send data.

Unsolicited grant service with activity detection (UGS/AD) is a combination of UGS and rtPS suitable for services that must transmit data quickly when it is available, in an environment in which data may not always be available. This would be used by a VoIP application that performs silence suppression. *Non-real-time polling service* (nrtPS) is similar to rtPS, but polling occurs at a lower rate. It is used for such tasks as bulk data transfer not requiring real-time transport.[10]

Classifiers are used to determine the type of packet being handled, for the purpose of assigning it to the correct service flow. Packets may be classified using any combination of header bits. These include the sending and receiving system MAC addresses (a unique permanent address assigned at the time of manufacture of a device) and the IP address (an address unique within at least a network, which may be assigned permanently or on an ad hoc basis). In addition, the DOCSIS system can classify on the universal datagram protocol (UDP) or transmission control protocol (TCP) port number or other characteristics as defined in Table 5.2 and earlier in this chapter.

Fragmentation is useful when handling relatively long packets. A packet may be broken into two or more fragments, which are sent individually, with other data interspersed, and then reassembled at the destination. This allows applications having smaller packets not to have to wait for large packets to be transmitted. This reduces jitter in time-critical applications such as VoIP.

Security

Table 5.2 also shows the security mechanisms built into the modem standards. Security requirements can be broken into several issues. The first issue is authentication. If you are going to send secured data to a modem, you want to know that the receiving device is a genuine modem that will do with the data what you want it to do. You need confidence that it is not a clone (imposter) that may look like a genuine modem but will do nefarious things with the data you send (such as eavesdropping on a neighbor's traffic). In DOCSIS 1.0, authentication consisted of simply getting the modem's MAC address and comparing it to known valid MAC addresses. If it is in the list, the modem is assumed to be good. This really offers no assurance, however, that someone has not built a pirate modem by cloning a legitimate modem's MAC address.

In DOCSIS 1.1, a far stronger authentication mechanism is included, the use of X.509 certificates. These certificates are described in detail in Chapter 22. In brief, the X.509 certificate is a unique electronic certificate stored in each modem and produced by a trusted third party called a certificate authority (CA). The certificate includes information that could have been known only to a legitimate modem, encrypted with the CA's private key. The CMTS knows the CA's public key, which it can use to decrypt the information in the certificate; if the information matches, the modem can be assumed to be legitimate.

Key exchange is the process by which the CMTS and the modem agree on a secret key (the data exchange key) to be used in encrypting data. The process used in DOCSIS 1.0 permitted the data exchange key to be transmitted, albeit encrypted using another encryption algorithm. This is not bad, but if someone were to learn the key exchange encryption key, then the key would be discoverable. In DOCSIS 1.1 and 2.0, the key exchange key is never sent from one side to the other, reducing the possibility that the key exchange key can be discovered.

Packet data unit (PDU) *encryption* is the same in all versions of DOCSIS, but in 1.1 and 2.0 each service flow can have its own encryption key.

5.2.6 More on QoS

Here we discuss the QoS mechanisms of DOCSIS 1.1 and 2.0 in additional detail. Unfortunately, space does not permit a full exposition of the features available, so we shall just provide a feel for what is going on.

Service Flows

Figure 5.8 illustrates several service flows to two different modems. Suppose cable modem 1 is handling two sessions: a voice call and streaming audio. The voice call is critical, in terms of both jitter and delay. The streaming audio requires somewhat more bandwidth. Because it is more tolerant of delay, its decoder can buffer a long string if necessary, making jitter less important.

Thus, the two services are assigned to different service flows. Cable modem 2 is handling a video conference, which requires more bandwidth and is not very tolerant of delay.

Service flows are set up for both modems. Separate service flows are set up in each direction as required, though we show only the downstream direction here. Note that there must be adequate bandwidth in the downstream channel for all service flows, because their bandwidth requirements add. (Figure 5.8 is more of a logical diagram than a physical diagram. Of course, all packets actually enter each modem, and the modem selects only those packets destined for it.)

There are two kinds of service flows: static and dynamic. Static service flows are set up during initial modem registration (when the modem first communicates with the CMTS). At least one service flow in each direction must be established at registration. This is known as a *primary* service flow and can be used for housekeeping and also for best-effort data communications (recall that *best effort* does not imply high-quality communications; quite the contrary, *best effort* means "if and when I can"). Dynamic service flows are established as needed during communications, as sessions are set up and torn down without rebooting the cable modem.

Service flows are identified by a 32-bit number called a *service flow identifier* (SFID) assigned by the CMTS. A service flow may exist even when it is not carrying any traffic. A service flow is said to be *admitted* if the CMTS and possibly the modem are reserving resources. It is said to be *active* if it is actually carrying data. If the service flow is admitted or active it also has a *service identifier* (SID) that identifies the traffic in the upstream direction to the CMTS.

Each packet entering the modem is identified by certain characteristics, described later, and assigned to the appropriate service flow by a *classifier*,

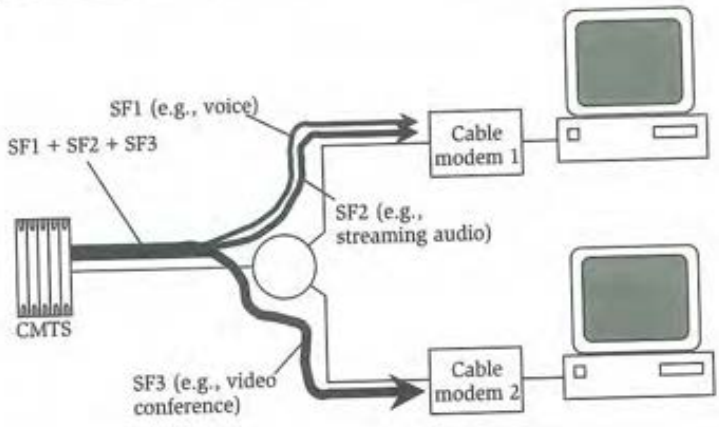


Figure 5.8 Downstream service flows.

a set of criteria used to identify the packet and associate it with a specific service flow. The service flow parameters will determine how to treat the packet. Thus, each service flow is handled in accordance with the policies set up for that service flow. In order to classify the packet, the classifier looks at the header on the packet.

As explained earlier in this chapter, the header comprises a number of bits added to the packet at its beginning. When the packet is placed into (usually) a TCP or UDP (layer 4) packet, a header is added. This header includes source and destination port numbers, which generally identify the program or process using the packet, as explained later in the chapter. When the TCP or UDP packet is placed into an IP packet, a header is added. This header can identify the hardware or portion of hardware using the packet. The link level control (LLC) also adds a header. These headers may be read by the classifier to determine the proper service flow to which the packet should be assigned. The headers can identify the client (end machine or program) sending or receiving the packet as well as identifying the type of packet. From this information, the classifier should have adequate information to identify the service flow, which, in turn, will know how the packet should be handled. If no classification exists for that packet, it is assigned to the primary service flow (the "best-effort" service flow).

An Example Dynamic Service Flow

A dynamic service flow must be set up when needed. To illustrate dynamic service flows, which are set up for one particular session, we use a streaming media example. Suppose the subscriber is downloading a song played in real time or an Internet radio station, something that requires relatively consistent packet delivery. That session is to be assigned its own dynamic service flow. Note in this example that we assume cooperation between the cable operator and the content-on-demand (COD) provider. It also assumes that the client software (on the subscriber's computer) has been customized for this application (or a Web browser is communicating with a COD application on the Web server). If these criteria are not satisfied, the subscriber can still download the streaming media but will have to do it as a best-effort stream, not in a service flow that guarantees the needed bandwidth and jitter.

Figure 5.9 illustrates setup of the streaming media session in a dynamic service flow. The service flow is *dynamic* because it is set up for one session and will be deleted ("torn down") at the end of that session. We are showing here a content server, perhaps owned by the COD vendor, which sends its content upon request to a cache server located in the cable operator's domain.

An optional cache server reduces the number of times the same file must be sent if more than one subscriber wants the same material at different times. The cache server will send the requested file to the appropriate CMTS (through the appropriate routers, using IP to deliver the file to the right point). The CMTS in

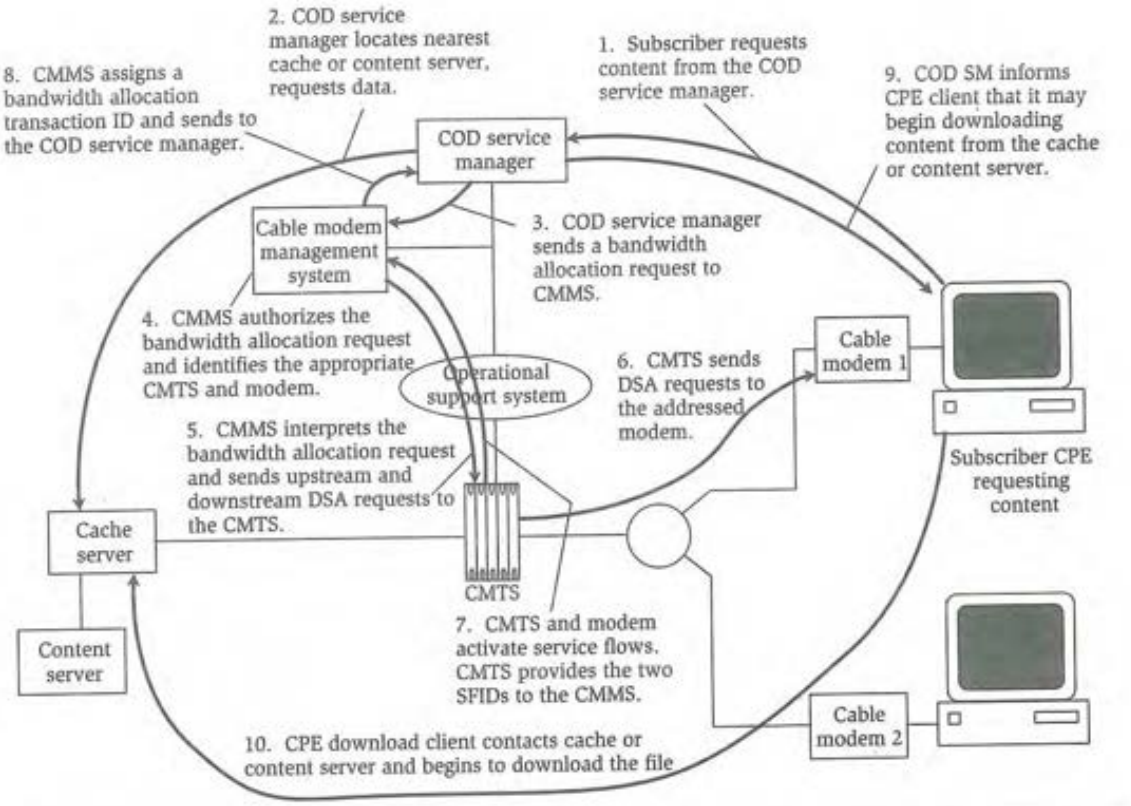


Figure 5.9 Streaming media setup.

turn sends the file to the subscriber's cable modem, which in turn delivers it to the computer.

Also shown are two functions considered part of the operational support system (OSS) for modems in the operator's domain. These are a COD service manager, which contains the relevant policies for offering services such as dynamic service flow, and the cable modem management system (CMMS), which keeps track of what CMTS and modem provide service to which subscriber.

In step 1, the subscriber requests content from the COD service manager, which locates the content on the cache server, if that is where it is, or on the content server. In step 2, the COD service manager requests data from the cache server on behalf of the subscriber (the "client"). In step 3, the COD service manager sends a bandwidth allocation request to the CMMS. In step 4, the CMMS receives the bandwidth allocation request, authorizes it, and identifies

the appropriate CMTS and modem. In step 5, the CMMS interprets the bidirectional bandwidth allocation request and sends a pair of upstream and downstream DSA requests to the proper CMTS, which, in step 6, sends the DSA requests to the addressed cable modem. DSA is one of three types of dynamic service commands used, the other two being dynamic service change (DSC) and dynamic service deletion (DSD). Each of these three commands has a request, a response, and an acknowledgment command. Collectively they are known as DSx transactions. The functions of each are obvious from the name. To add a dynamic service flow, a DSA request is sent, which is acknowledged by a DSA response (usually granting the request), followed by a DSA acknowledgment. Similarly, DSC changes a dynamic service flow and DSD deletes it.

In step 7, the CMTS and modem activate the service flows, and the CMTS provides the two (upstream and downstream) service flow identifiers (SFIDs) to the CMMS. In step 8, the CMMS assigns a bandwidth allocation transaction identification and sends it to the COD service manager. At that point (step 9) the COD service manager informs the CPE client (the software being used by the subscriber) that it may begin downloading content from the cache, if that is where the file is located. Finally, in step 10, the CPE client is now ready to contact the cache server and to begin to download the file. Note that the classifiers installed in the CMTS and the modem for these COD flows recognize the download traffic between the COD cache server and the client and then the selected service flows provide the requested QoS to these packets. Meanwhile, other traffic is managed by the other service flows and is not given the special QoS treatment afforded the COD traffic.

At the conclusion of the session, a similar set of processes is used to delete the service flow. When the COD cache server indicates the download has completed, the COD service manager uses the bandwith allocation transaction ID to inform the CMMS which bandwidth resources to release. The CMMS in turn relays a pair of DSD requests to the CMTS to remove the COD service flows from the modem. Note that if a subscriber attempts to download material using download client software that does not support the request for dynamic service flow, the subscriber can still do so. However, he or she will not receive a dynamic service flow and is at the mercy of how busy the system is to set the rate at which material is downloaded.[11]

5.3 Non-DOCSIS Protocols

This section deals with a number of protocols of importance to cable television technologists. Of necessity, this is no more than a quick-read introduction for those needing basic familiarity with the protocols; it will not make you an expert. Books have been written about each protocol, and we shall reference some of them for those needing more information.

5.3.1 Asynchronous Transfer Mode (ATM)

Asynchronous transfer mode (ATM) is a connection-oriented layer 2 protocol that originated in the telephone industry to provide transport for multiple voice conversations and other data. It features fixed-length cells (as opposed to variable-length packets — see Section 5.3.2 on Ethernet). As shown in Figure 5.10(a), the cells are 53 bytes long, with the first 5 bytes being the destination header. Note that in ATM we refer to fixed-length "cells" rather than the variable-length "packets" of Ethernet and IP. ATM systems operate continuously, with one cell followed immediately by the next, even when there is no data to transmit. Typically, when a connection is set up, sufficient bandwidth is reserved for that connection, and it is reserved until the connection is torn down, regardless of whether it is carrying traffic.

ATM is sometimes used to interconnect hubs serving an HFC plant, but due to cost it is not popular in subscriber service. (Certain concepts of ATM do survive in a number of subscriber-facing services, though.) ATM is known for low jitter (cells arrive at a regular interval) and excellent quality of service. A number of mechanisms are provided to ensure reliable delivery of cells.

As shown in Figure 5.10(b), information delivered to an ATM system interfaces via the *ATM adaptation layer* (AAL), which has two functions. Since many packets from other protocols, such as Ethernet, can be much longer than the 48 data bytes of ATM, these longer packets must be divided into segments at the transmit end and then reassembled into longer packets at the receiver. This is done by the *segmentation and reassembly* (SAR) function.

The other AAL function is performed by the *convergence sublayer* (CS). It provides interface to different types of data packets, called *service data units* (SDUs). Different SDUs require different types of processing. For example, you might be handling video in constant or variable-bit-rate streams. These need a

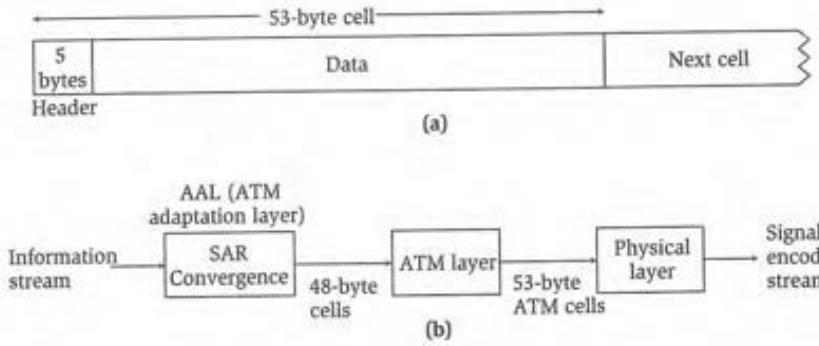


Figure 5.10 (a) ATM cell stream. (b) Building the ATM cell.

different interface, as does an IP packet or other types of data. Five types of CS interfaces are specified.

AAL1 is for traffic generated at a constant bit rate, which must be delivered at that same data rate (after a fixed delay). Intended applications are voice and constant-bit-rate video or audio.

AAL2 is for traffic generated at a variable bit rate. A time stamp is used to control delivery of the packets at the receiver. The SAR must segment variable-length packets generated by the CS. The source and destination must agree on how to deal with timing and cell loss, as in AAL1.

AAL3 is suitable for a connection that occupies one or more variable-sized packets. Two options are provided for error quality. In one option, the CS guarantees error-free delivery. If an error occurs, the destination CS requests retransmission. In the other option, no guarantees are made for error-free delivery.

AAL4 is similar to AAL3 except that it emulates connectionless (datagram) service. The CS layer generates a single variable-sized packet, which is segmented into one or more cells by the SAR sublayer.

AAL5 is for IP packets. The IP packet is packaged into a CS packet that contains a length indicator (for the IP packet) and a CRC calculated over the complete CS packet. The ATM cell header indicates whether or not the cell is the last one of the CS packet.[12]

5.3.2 Ethernet

No layer 2 networking standard is more prevalent today than Ethernet. This connectionless protocol has dominated office networking almost since the beginning of networking, and a large number of homes being built now have wiring in them to enable Ethernet networks. Ethernet was developed at the Xerox Palo Alto Research Center, PARC, by Bob Metcalf in the early 1970s. Metcalf's work built on that of Norman Abramson and his colleagues at the University of Hawaii in the late 1960s. This team developed a network of data communications via radio between various Hawaiian Islands and ships plying the waters around them. Appropriately, the protocol they developed was and is known as *Aloha*.

The pure Aloha protocol was simple: When a station had something to say, it simply transmitted its information and waited for an acknowledgment by the receiving station. If the acknowledgment didn't come in a reasonable length of time, the sending station assumed it didn't get through because another station transmitted at the same time. A *collision* was said to have occurred. It and the other station (which would not have been acknowledged either) waited different random times and then retransmitted.

In any radio communications channel, you cannot assume that a station wishing to transmit can hear another station that is already transmitting, so this

was not required in the Aloha system. Channel utilization is low, because as more stations try to send more data, the probability of a collision increases. Pure Aloha systems can achieve maximum channel utilization of about 18% due to this problem. By coordinating time slots in which stations can transmit, a *slotted Aloha* system is derived, which has a theoretical utilization of about 37%.

In wired networks such as Ethernet, you can assume that one station can hear another, and this allows improvements. Metcalf added a mechanism that listened to a channel before it tried to transmit (*carrier sense*). This should resolve most collisions. But it is still possible for two stations to start transmitting at very nearly the same time, so after a station starts transmitting, it listens to its own outgoing signal to see if a collision occurs (*collision detection*). When a collision is detected, the station stops transmitting and waits a random time before trying to transmit again. Multiple stations can access this network, so we have *carrier sense, multiple access with collision detection*, or a CSMA/CD system.

The Ethernet standard was published in 1980 by a consortium of three companies, Digital Equipment Corporation (DEC), Intel, and IBM. This standard has become known as the DIX standard, after the three companies. About the same time, the IEEE[13] started to document the standard, and continues today to be the custodian of the Ethernet standard, known as IEEE 802.3.* (There are other communications standards in the 802 family — perhaps one of the best known today is the wireless 802.11 set of standards.) The original DIX standard (which reached version 2.0) and the 802.3 standard are similar, though there are a few differences. We shall describe the IEEE version of Ethernet.

There have been a number of different physical layer implementations of Ethernet; here we discuss only the most commonly encountered implementations. Ethernet physical interfaces are described by a number indicating the speed, followed by the word *Base*, followed by a designation indicating the medium over which the signal is transmitted. 10BaseT indicates a 10-Mb/s transmission speed over twisted pair cable. 100BaseTX is a 100-Mb/s signal transmitted over the same types of twisted pair cable. These two form the basis for the most common Ethernet interfaces in use today. Much equipment can interoperate between the two speeds; this is known as a 10/100BaseT interface. Other popular forms of Ethernet operate at a date speed of 1 Gb/s: 1000BaseT (twisted pair), 1000BaseSX (short-distance optical using multimode fiber), and 1000BaseLX (long-distance optical using single-mode fiber). 10-GB Ethernet has been standardized.

10/100BaseT

There is almost zero probability that your office is *not* using 10/100BaseT for its internal network. The numerals before the word *Base* represent the signaling

* The 802.3 standard does not use the term *Ethernet*, because the IEEE is properly concerned about being tied to any commercial terms.

speed, either 10 or 100 Mb/s. The *T* indicates twisted-pair transmission medium. Technically, there should be a dash between *Base* and *T*, but we follow the common practice of omitting the dash. There are several variations of 100BaseT communications. The one in common use is really 100BaseTX, but since the others are not used, often the *X* is omitted.

Many homes built in this century are wired with cable to support 10/100BaseT data transmission. The medium is two twisted pairs, one for each direction of communication. The gage of wire, the insulation, and the twist are selected so that the differential characteristic impedance is 100 ohms. Characteristic impedance has exactly the same definition as in the 75-ohm coaxial cable you are used to, except that the transmission is balanced with respect to ground rather than single-ended.

Most of the cable used today is unshielded twisted pair, designated UTP. You may on rare occasion see shielded twisted pair STP. The standard cable consists of four twisted pairs, of which two are used for one Ethernet connection. There are varieties of Ethernet that use all four pairs, but you are not likely to encounter them until you get to gigabit Ethernet. While emphatically *not* recommended, sometimes the two other twisted pairs are used for telephone connections. Due to crosstalk potential, it is *strongly* recommended to use separate cables for voice and data, in both residential and commercial wiring. The connector is the RJ-45, which looks like a larger version of the RJ-11 connector used for telephones. (RJ stands for *registered jack*, a U.S. telephone industry term.)

Manchester encoding is used to provide a clock signal and to remove any dc component of the data. Manchester encoding is described in Section 4.2.2. The voltage on each pair is balanced, with a minimum amplitude at the transmitter of 900 mV pin to pin (p–p) and a maximum voltage of 2.63 V p–p at the crest of an overshoot. The receiver works with as little as 160 mV p–p. Contrast this with the ringing voltage of a telephone pair, which can be as high as 108 V, and you see why it is not a good idea to run telephone in the same cable as Ethernet. Unfortunately, some home builders are wiring both phone and Ethernet connections in the same cable, despite an abundance of wisdom that says not to do it.[14] Laboratory tests done by the writer in 2003 indicated that it will sometimes work, but in no way can we endorse this reckless practice.

1000BaseT

Gigabit Ethernet was standardized in 1999. There is one standard for twisted-pair cable and two for optical cable. 1000BaseT is the reference for the twisted pair standard, which uses all four conductors of a Cat5 or better cable (see later). The signal encoding, called 4D-PAM5, consists of four symbols, each of which contains two bits. All four wire pair are used to transmit signals in each direction, so 8 bits are transmitted at a time. A *hybrid* is used to separate transmission in each direction, similar to what is done in conventional analog telephony. Hybrids may be built with transformers or with operational amplifiers. As with the slower speeds, balanced transmission is used, with a voltage range of ±1 V p–p.

1000Base-X (Fiber Optics)

There are two forms of gigabit Ethernet over fiber-optic cable in use. 1000Base-LX is a long-wavelength (usually 1310 or 1550 nm) signal using single-mode fiber just as is used in HFC plant. It is intended for long-distance operation, with commercially available transceivers (often called *GBICs*) for distances exceeding 70 km. 100Base-SX is a short-wavelength, short-distance transmission standard that uses multimode fiber. Both standards use one fiber for each direction.

The signal uses 8B/10B encoding, in which a 10-bit symbol is substituted for every 8 bits of data. This does create 25% overhead — the actual transmission rate for gigabit Ethernet is 1250 Mb/s. It is used to ensure adequate transitions for clock recovery and no dc component. 8B/10B encoding is explained more fully in Chapter 19.

Twisted-Pair Wiring

The cable commonly used is called category 5, or simply *Cat5* cable. There have been several generations of cabling used for voice and data, with the number incremented for cables that exhibit better return loss and radiation characteristics. Besides Cat5, there is a Cat5e (*enhanced*) cable and a Cat6.

Figure 5.11 illustrates two standard ways to wire RJ-45 plugs. The Telephone Industries Association (TIA) standard T568 lists two wiring sequences, both shown in Figure 5.11. The "preferred" T568A standard was formerly used more often, but today the optional T568B wiring sequence seems to be more popular. The only difference is that wire pairs 2 and 3 are interchanged. Note that the wire pair numbering always stays with the same color of wire pair. Pairs 1 and 4, which are not used in the most common 10/100BaseT wiring, are the

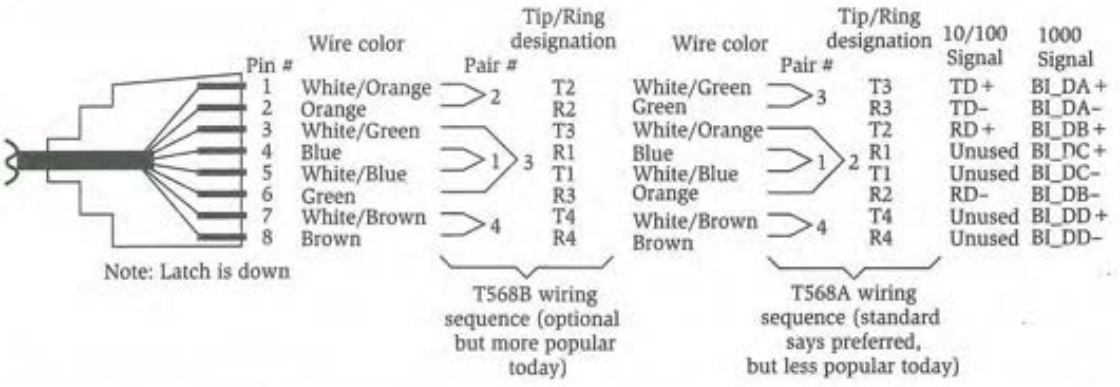


Figure 5.11 Common RJ-45 wiring.

same for both options, but wire pairs 2 and 3, which are used for the most common Ethernet wiring, are different in the two options.

The "signal" designations are for transmit data + and − and receive data + and −. For gigabit Ethernet on twisted pair (1000BaseT), different designations (standing for *bidirectional*) are used, since all pairs are used and signaling is transmitted in both directions simultaneously. Recall that these are balanced transmission lines, so the designation is to allow correct polarization, or phasing. The *tip/ring* designations go back to the early days of telephony, when 1/4-inch phone jacks, similar to the large jacks used in some audio equipment and for headphone connectors, were used in central offices (COs). The connection on the end of the plug was the "tip" and the other connection was the "ring." Note that the ring wire is always a solid color, and the tip is always white with a small trace of that same color.[15,16] (For stereo audio connections made on 1/4-inch phone jacks, a third connection is added called a *sleeve*.)

The important thing is that the two ends must be wired pin to pin on each end, and the twisted pairs must be followed. If you violate the twisted pair, for example, by interchanging the wires on pins 2 and 3 so that you still have pin-to-pin electrical continuity but don't have the transmit and receive pairs as separate twisted pairs, then the cable will probably not work, because you have upset the 100-ohm characteristic impedance, setting up standing waves (in the baseband data world this will manifest itself as pulse distortion). Furthermore, you will have excessive crosstalk between transmit and receive as a result of not having one signal (transmit or receive) on a twisted pair. A twisted pair driven with a balanced source and load does not radiate or pick up radiation, at least to the extent the system is truly balanced. This is the same as the issue described under Line Balance in Chapter 6.

Interconnecting Ethernet Devices

Figure 5.12 illustrates a basic in-home (or office) network with two computers and a game console. Other devices could be connected. All wiring is home-run from a hub or central device. The specified maximum distance between any endpoint and the hub is 100 meters. The end devices, computers, game consoles, or whatever are known as *data terminal equipment* (DTE). The central device has a number of names, depending on what it does and who is naming it.

Ethernet connections (at least the common ones you are likely to encounter) are never looped through from one device to another. Despite the attempts of some home installers to install loop-through connections, it doesn't work. *Loop-through* means that the same wire pairs are run from one outlet to another, after originating at a hub. Suppose you have a loop-through connection as shown on the game console in the figure. If the game console (or computer or whatever) is placed at the end of the loop-through, it will work. But if it is placed at the first point, as shown, the cable from there to the end of the loop-through will act as an unterminated stub and will reflect pulses, disrupting the performance of the

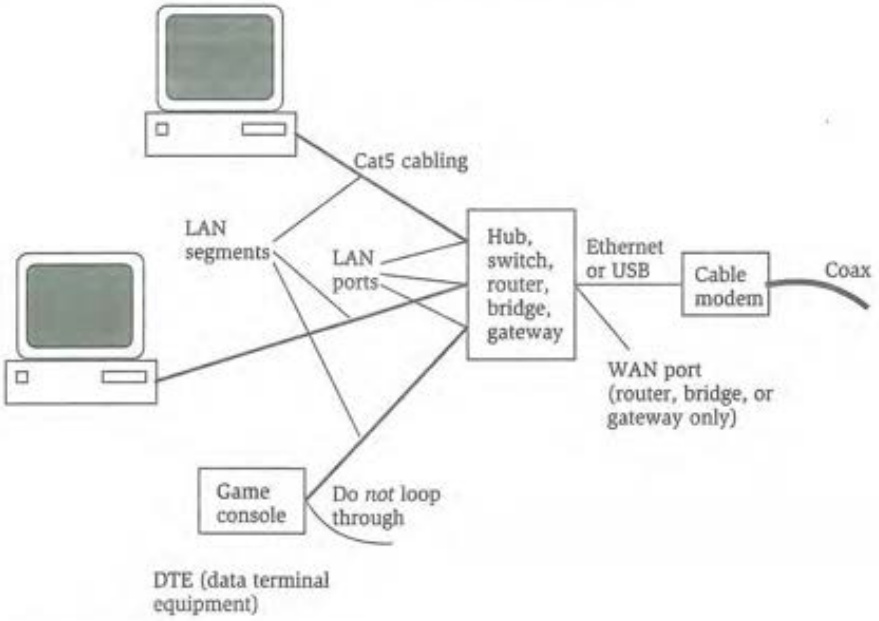


Figure 5.12 In-home network.

link. We have found that in some cases, terminating the end connection point will allow the system to work, but this is only an emergency solution, not a recommendation.

It is true that one Ethernet connection will listen for others transmitting before it transmits, but physically what happens is that there is a central hub that can echo signals entering one port to all other ports. The hub may perform other functions as well. A *hub* is a common connection point for devices in a network. Hubs are commonly used to connect segments of a LAN, as shown in Figure 5.12. When a packet arrives at one port, it is copied to the other ports so that all segments of the LAN can see all packets.

A *switch*, or sometimes *switching hub*, is a bit more intelligent. It filters and forwards packets between LAN segments. Each DTE is assigned a unique identification called a *MAC address*, described later. A switch learns the MAC address of each DTE connected to it; when it gets a packet with that MAC destination address, it forwards the packet only to the appropriate port. The switch learns the MAC address of each DTE by listening to what comes in on that port. Each transmission includes a destination MAC address and a source MAC address. The switch listens to the source addresses to learn what is on each port. This usually suffices to allow the switch to know how to route any packet it receives.

However, sometimes a packet will arrive with a destination MAC address the switch doesn't recognize. In that case, the switch will replicate the packet on all ports, just as a hub would do. When the correct port (if any) receives the packet and sends an acknowledgment, the switch will learn the correct port for that MAC address.

A *router* is a yet more intelligent device that can operate at layer 3 as well as layer 2. It often is placed between two LANs or between a LAN and a WAN. We shall cover their functions in more detail later.

A *gateway* is a router used to connect two or more networks, as shown in Figure 5.12. It may have a number of higher-layer services, which will be described below.

A *bridge* is a low-end gateway that interconnects two LANs or two segments of the same LAN.

Ethernet Frames

Figure 5.13 shows the makeup of an Ethernet *frame*. A frame holds one *packet* of data. The first 8 bytes are the preamble. Some Ethernet systems don't transmit continuously, so the preamble is used to synchronize a receive clock before data is transmitted. In other Ethernet systems, data transmission is continuous, but the preamble has been retained to keep the frame structure.

The next 6 bytes are the destination address — the MAC address of the intended recipient of the packet. Next is the source address, the MAC address of the sender. This is needed to allow the recipient to respond to the source and also to allow switches to learn the MAC address on each port, for efficient packet routing. Following the address fields is the type/length field, which is used for several purposes. It can identify the higher-layer network protocol carried in the data field, as described next. It also identifies the length of the packet. Unlike ATM cells, Ethernet packets are of variable length, so something must tell the receiver when a packet ends.

Next is the data, the thing we want to transmit. It may have headers for higher-layer protocols embedded in it. The length of the data is between 46 and 1,500 bytes. Following the data is the frame check sequence, a cyclical

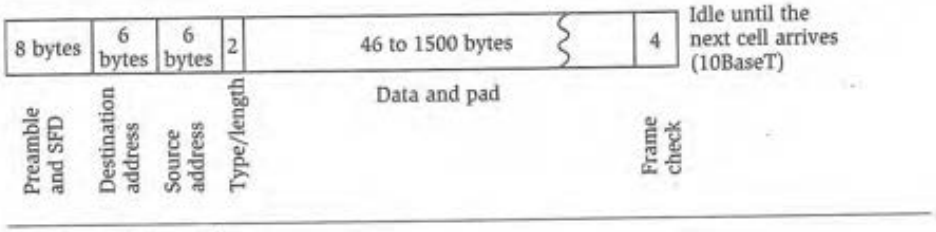


Figure 5.13 Ethernet frame.

redundancy code used for error checking but not correction. Error checking is described briefly in Chapter 3.

Ethernet Addressing

Every Ethernet port in every piece of equipment is assigned its own unique address, called the MAC (media access control, a layer 2 feature) address. Each manufacturer applies for and receives a range of MAC addresses, which are assigned individually to each and every piece of hardware manufactured. The MAC address is 48 bits (8 bytes) long, so the total available address space is about $2^{48} = 281.5 \times 10^{12}$ addresses. The first 24 of those bits are called the *organizationally unique identifier* (OUI). They identify the organization that builds the interface. The IEEE Standards Association administers this portion of the address field.

This assigned address is a *unicast* address, meaning that it is used when another device wants to contact that Ethernet port and only that Ethernet port. When a device starts to receive a new packet, it listens to the destination MAC address. If the transmitted address matches its address, it receives the packet. Otherwise, it discards it. However, there are times when you need to communicate with multiple devices. You can use *multicast* addresses for this. Any number of Ethernet ports may be set to listen to the same multicast address and to receive the data. Finally, there is a *broadcast* address (all 48 bits are 1s), used to communicate with all devices on the network.[17]

Ethernet addresses are written in a manner intended to make it easy to deal with a 48-bit binary number. They are written as pairs of hexadecimal numbers, separated by dashes. A *hexadecimal* number is a base-16 number represented as shown in Table 5.3. Each byte (8 bits) of the MAC address is written as two hexadecimal numbers, separated by a dash. For example, the Ethernet card in the writer's computer, which interfaces to the in-home network, has the MAC address 00-03-6D-15-C0-97. Note that there are six sets of two hexadecimal numerals, each set representing 8 bits, for a total of 48 bits.

The card that handles the interface between the computer and the network is called the *network interface card* (NIC). Even if the network interface is built into the computer or other equipment, it is still called an NIC.

5.3.3 Internet Protocol (IP)

Internet Protocol (IP) is not just the Internet. It is a very important connectionless layer 3 protocol used to move packets around on the Internet as well as on a lot of other networks. IP allows any computer on the Internet (or another IP network) to locate and exchange packets with any other computer. You often see TCP/IP together, where TCP is a layer 4 protocol often used with IP (as in, for instance, UDP). TCP and UDP are discussed later.

Table 5.3 Hexadecimal Numbering

Base 10	*Base 2*	*Hex*
0	0000	0
1	0001	1
2	0010	2
3	0011	3
4	0100	4
5	0101	5
6	0110	6
7	0111	7
8	1000	8
9	1001	9
10	1010	A
11	1011	B
12	1100	C
13	1101	D
14	1110	E
15	1111	F

IP addresses are expressed in the *dotted decimal* format. In version 4 of IP, addresses are 32 bits long. Table 5.4 illustrates the dotted decimal address. The top row gives the dotted decimal representation; the second row expresses this as individual decimal numbers, the third row shows the corresponding binary numbers. This IP address was assigned to the WAN port on the writer's gateway connected to his cable modem at the time of writing. The assignment came from his cable operator. As shown later, the IP address may change from time to time.

The contents of an IP packet are called a *datagram*. Each datagram, as it traverses the Internet, retains the same source and destination IP addresses, even as it traverses different layer 2 networks. For example, a packet may travel part of the distance on a token ring network, partway on ATM, and partway on Ethernet. These are layer 2 protocols, and each will put on its own layer 2 header as the IP datagram enters that network and strip its layer 2 header when it is handed to another network. But the IP header stays with the datagram during the whole process (though it will be modified in each router through which it passes), defining the ultimate destination of the datagram. (An exception is that

Table 5.4 Dotted Decimal Notation for IPv4 Addresses

Dotted decimal		*24.98.160.78*		
Base 10	24	98	160	78
Base 2	00011000	01100010	10100000	1001110

a firewall between networks or between a network and a computer may substitute its own IP address for that of the ultimate process using the datagram.)

When switching in a network is done by looking at the layer 3 (IP) address, the device doing the switching is usually called a *router*. Routers technically have hosts built in, because they themselves can be addressed (they have their own IP addresses), for management purposes.

IPv4 Versus IPv6

IP version 4 (IPv4) is in common use today. For years there has been a desire to replace IPv4 with IPv6, but the difficulty of managing such a change has prevented widespread deployment of IPv6. There are several new features in IPv6, the most notable being the expansion from a 32-bit address to a 128-bit address. This will simplify address management, removing the need for assigning temporary private IP addresses to devices. In addition, the header format is streamlined. A flow label is added to denote a stream of traffic with special QoS needs. Support for both authentication and confidentiality are mandatory, and there are a few other changes. Unless otherwise noted, we are describing IPv4 in this book.

IP Embedded in Ethernet

Figure 5.14 illustrates an IP datagram inside an Ethernet frame. This figure is laid out with four bytes (32 bits) in a horizontal row. Each horizontal row follows serially the one above it.

Compare Figure 5.14 with the illustration of an Ethernet frame in Figure 5.13, which identified all the elements of the Ethernet frame consistent with Figure 5.14. The "Data and pad" of Figure 5.13 are expanded in Figure 5.14 to show the elements of the IP datagram. The elements of the IP header are summarized next.

The IP Header

The first four bits, *vers*, contain the IP version number, currently 4 (0100) and possibly going to 6 (1010) at some point. The next four bits, *hlen*, counts the number of 32-bit units in the IP header. The basic IP header as shown is 20 bytes, or five units of 32 bits. We show later that some IP headers are longer as a result

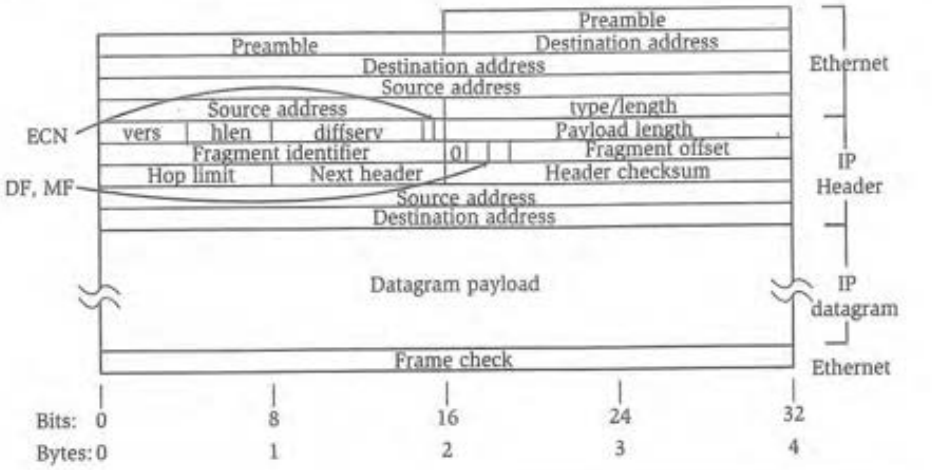


Figure 5.14 IP Datagram Inside Ethernet Frame.

of options being embedded. IP defines a number of options that may create a longer header. These are covered in the endnotes. Options include, but are not limited to, the ability to record the route taken by the datagram, the ability to force a particular route, and time stamping.

The next field is a 6-bit field labeled *diffserv*, for differentiated service. The 6 bits contain a so-called *diffserv codepoint* (DSCP) value, which tells a router how to handle the packet. Special treatment may include expedited forwarding or extra protection from loss. Most values for this field are not defined in IP but rather are used by different networks in accordance with each network's policies. This field is an important enhancement to IP to bring it closer to parity with the QoS capabilities of ATM or similar protocols. It allows datagrams to be given higher priority if the datagram is something, such as voice, that needs to get through a network quickly. (Note, however, that just because ATM has good QoS features does not mean that they are always used.)

The final two bits in this field, ECN, are the *explicit congestion notification field*, which provide cooperating routers with a way to improve management of congestion on a link. When a router handling the packet notices that a link is getting congested, it sets these two bits to 11 before forwarding the packet on the congested link. When the packet is ultimately received, the receiving host sees that one of the routes the packet took is congested, so it slows down the sender, using methods shown later in the section on Transmission Control Protocol.

The next two bytes are the *payload length*, which defines the length of the datagram, not counting the header. IP datagrams can be quite long, though they may well have to be segmented into several packets or cells (depending on the protocol on which it is carried) on any particular transport segment. For

example, if the datagram is being carried on Ethernet, as shown in Figure 5.14, the maximum amount of data in the Ethernet frame is 1500 bytes (this includes the IP header and the datagram). If the datagram goes through an ATM network, it will have to be segmented into cells no longer than 48 bytes, as described earlier. If the datagram is too long to fit, it will have to be segmented into several packets. The *fragment identifier* field is added at the sender, to be used to uniquely identify the packet to which a fragment belongs so that the fragments can be reassembled at the destination. This field is set to a value even if the sender is not fragmenting the packet, because it is possible that a later network will have to fragment it.

The next three bits are flags that instruct routers on segmentation. The first bit is always 0. The next bit, DF, indicates that the packet may not be fragmented. If a router encounters a datagram that is too long and the DF bit is set, then the router has no recourse but to discard the packet. The next bit, MF, indicates that more fragments follow. It is set except on the last segment of a fragmented datagram. Finally, the *fragment offset* field specifies how far into the datagram the fragment is so that the fragments can be reassembled as one datagram by the time they reach their destination (if not before). It is quite possible that the segments will arrive out of order, so segments received before predecessor segments will have to be held until all segments are received.

The next field is the *hop limit*. (This was originally called the *time to live* (TTL) field). The sender sets this field to some value; every time the datagram passes through a router, the field is decremented by 1. If the value reaches 0, the packet is discarded. The hop limit protects the Internet against an undeliverable packet, which would otherwise hop around a network forever if it could not be delivered. This can happen if a *routing loop* is created; that is, an error in a routing table sends a packet back to the previous router that handled it, which sends it to the router with the errored routing table, which sends it back, etc. If nothing caused the packet to be discarded, it could bounce around forever in the Internet, tying up resources needed by packets that can be delivered.

Next is the *next header* field, also called the *protocol field*. This indicates the next higher header, which may be a layer 4 header or may be intimately related to IP. We shall discuss some of these protocols later. Finally, the *header checksum* provides protection against corruption of the header. Each router that handles the datagram computes the checksum by doing a 16-bit addition of the header, discarding carries. The receiving router does the same computation; if the two disagree, it means that an error was made in the header, so the entire packet is discarded.

Finally, the IP header includes the source and destination IP addresses. These don't change as the datagram traverses the network. Each router handling the packet reads the destination address to know where to send the packet (controlled by a massive and continuously changing routing table in the router) and reads the source address to update its routing table for that address.

Types of IP Addresses

There are three types of IP addresses: unicast, multicast, and broadcast. *Unicast* addresses are the most straightforward. They are assigned to identify a single interface (for most practical purposes this equates to a single host, though there is a difference). A unicast message is intended for one and only one interface. *Multicast* addresses identify a set of interfaces. This allows a message to be generated once and sent to a number of interfaces. A router with receiving hosts attached to two or more ports will replicate the packets on each port. Multicast addressing is useful for video conferencing, in which the same signal is to be sent to several participants. Similarly, it is used in IP distribution of video (IPTV). If one router needs to pass messages to several other routers, it can do so using multicast addresses. *Broadcast* addresses are a special case of multicast addresses. They identify all interfaces on a network. The use of broadcast addresses is discouraged.

Internet Control Message Protocol (ICMP)

Internet Control Message Protocol (ICMP) is a layer 3 (network) protocol, but it sends its messages inside IP datagrams. The *next header* field (see Figure 5.14) identifies ICMP. ICMP coordinates many aspects of the operation of a network, including controlling address assignments, reporting errors, and providing diagnostic support.

Router discovery is one important function it performs. This allows hosts (any device on a network) to discover what routers are connected to the network. Periodically, routers send *router advertisement* messages to all systems on the network. The router advertisement message includes the address of every router known to the sending router and a relative preference value. When possible, all hosts should select the router with the highest preference value. The message also includes a lifetime field. If a host doesn't hear from a router within that lifetime (typically 30 minutes), it must assume that the router is no longer available. The router advertisement messages are sent more frequently than this.

Just knowing the addresses of all routers on a network does not ensure that a host will direct a message to the router that is the best path to the destination. It selects a router to which to send the message. If it selects the wrong router, that router redirects the message to the correct router and also sends an *ICMP redirect* message to the host, saying, in essence, "Next time you want to send a message to this host, send it to this other router, not to me."

Internet Group Management Protocol (IGMP)

Like ICMP, *Internet Group Management Protocol* (IGMP) is a protocol in its own right, but it is an integral part of any IP implementation. Hosts use IGMP to announce (and later renounce) their membership in groups. Routers listen to

these messages to track group membership. They then know how to forward datagrams addressed to groups.[18]

Dynamic Host Configuration Protocol (DHCP)

IP addresses are assigned to individual hosts, that is, all computers on a network. These IP addresses may be taken from public IP addresses, or, since IPv4 has a limited address space, most machines that don't have need for a public address are assigned private IP addresses owned by the network to which the host is attached. This allows reusing addresses on the Internet and simplifies the management of the network. IP addresses may be assigned manually to hosts, or they may be assigned automatically. Automatic assignment is usually preferable when possible.

The protocol used to assign IP addresses is the *Dynamic Host Configuration Protocol* (DHCP). When a host, such as your PC, comes up on a network, it asks for an IP address if it has not been assigned a permanent one. The host sends a DHCPDISCOVER packet. Connected somewhere to the network is at least one *DHCP server*, whose job it is to assign IP addresses to hosts. When a DHCP server receives a DHCPDISCOVER packet, it returns a DHCPOFFER packet containing an IP address and other information. It is possible for more than one DHCP server to be resident on a network, in which case the host receives more than one offer. It selects one and sends a DCHPREQUEST message to that server. Assuming the offer is still valid (for example, the IP address has not been assigned to another host in the meantime), the DHCP server returns a DHCPACK (DHCP acknowledgment) message.

Once the host receives the DHCPACK message, that IP address belongs to it for the duration of the *lease*, that is, the time for which the DHCP server has said that the address is valid. Before the end of a lease, the host must request another IP address by starting the process again. The lease allows an IP address to be placed back in the pool of available addresses if the host goes away for any reason, such as a power-down.[19]

5.3.4 Layer 4 Transport Protocols

Most Internet traffic is handled using one of two layer 4 protocols: TCP or UDP. Layer 4 protocols have two primary responsibilities. First, they must distinguish between different applications running on a system. On the same computer, you can download email, stream audio, and open a Web page all at the same time, over the same physical connection. It is the layer 4 protocols that are making sure packets are directed to the correct application.

The second responsibility of layer 4 protocols is reliability. IP is a best-effort service, which does not guarantee delivery of packets and does not guarantee that the packets delivered will be error free. Layer 4 protocols either improve the reliability of delivery slightly (UDP) or they improve it greatly (TCP).

Ports

In order to distinguish between applications, the layer 4 protocol uses a *port number*. Data is transferred between a communications system and an application using a port number. A good way to think of a port is to recognize that every application getting data from a network must somehow transfer that data in the computer, from the program that sends and receives data to the application program. For example, suppose you are streaming audio. IP is running in your computer to receive datagrams bearing the streaming audio. Another application is running to play the audio. The two must exchange data. The application playing the audio asks for a port number when it starts. The operating system assigns a memory address corresponding to that port number. Data from the communications program is placed in that address and then picked up by the streaming audio program. The address performs the function of a mailbox.

When an application requests data, for example, by your clicking a hot key for a streaming audio site, the application sends a request to the server having the data, and in the request messages it includes the port number for the return datagram. The server puts this port number in the return datagram. The communications software was informed by the operating system of the address corresponding to the port number, so when it gets a datagram with that port number, it knows where to store the datagram in memory. The streaming software then reads that address and processes the datagram into audio.

Some port numbers are created on an ad hoc basis as they are needed; others are standardized and predictable. The standardized port numbers are called *well-known ports*. Table 5.5 lists a few of the well-known ports.

Sockets

Ports identify application protocols, but they don't identify application programs running in real machines. To do so requires not only a port number but also the address of the machine running the program. This combination of a port number and an address is known as a *socket*.

Table 5.5 Examples of Well-Known Ports

Port(s)	*Transport(s)*	*Application*
20, 21	TCP	File Transfer Protocol: used to transfer files between computers
25	TCP	Simple Mail Transfer Protocol: for email transfer
53	TCP, UDP	Domain name service: maps a URL to a server's IP address

User Datagram Protocol (UDP)

User Datagram Protocol (UDP) is a layer 4 protocol used to deliver datagrams not requiring or not being able to use some of the features of TCP, discussed shortly. UDP is a simple, connectionless protocol that is the layer 4 protocol of choice for such applications as streaming audio and video and VoIP (voice on Internet Protocol—see Chapter 6) packet delivery. Applications using UDP are frequently characterized by the need for rapid (low-latency) delivery, where it is not feasible to correct for transmission errors. UDP is also used where a transmission is bound for more than one receiver, making acknowledgment moot.

For example, suppose you are handling a VoIP voice transmission. If a packet is received in error, there is not much you can do about it anyway. You could ask for retransmission, but by the time the request goes to the sending machine and it retransmits the packet, it is much too late to use that packet. So you do the best you can with the packet received in error. This might mean that you discard it and replace it with the previously received good packet. Similarly, you don't have the luxury of reordering packets received out of order, because this will require time, and time is of the essence in handling audio packets. Two other protocols often used with UDP are the *Real-Time Protocol* (RTP) and *Real-Time Control Protocol* (RTCP), both described in Chapter 6.

UDP headers and data reside in the datagram payload of Figure 5.14. The UDP header is simple, consisting of just four parts: the source port, destination port, datagram length, and checksum. Following the header is the application data, the desired information to be transmitted.

Transmission Control Protocol (TCP)

Transmission Control Protocol (TCP) is a more sophisticated, connection-oriented protocol used for reliable delivery of data that is not required to be delivered in real time. TCP can correct errors in transmission. It can detect packets received out of order and put them back in the correct order. It eliminates duplication of packets and requests missing packets.

Figure 5.15 illustrates a TCP message in an IP datagram, in an Ethernet packet. Remember that the Ethernet encapsulation may or may not be present, depending on what layer 2 network is carrying the datagram. The first two elements of the header are the source and destination ports, the purpose of which was explained earlier.

The *sequence number* is the field that allows packets received out of order to be put back in the correct order. The sender assigns a sequential number to each byte of the application data. The sequence number is the sequential number of the first byte of the TCP packet. When the receiving application receives a packet, it looks at the sequence number and compares it with the expected sequence number, computed by adding the number of bytes in the previous packet to the sequence number of that previous packet. If the sequence number is the one expected, then the receiver knows that it can simply assemble the data

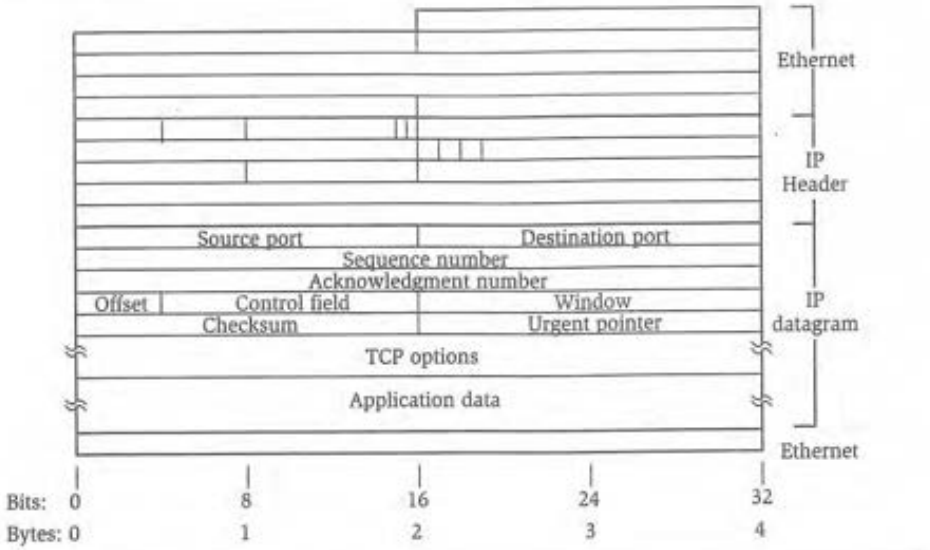


Figure 5.15 TCP in IP in Ethernet.

in the two packets to pass it to the next higher layer. If a sequence number is duplicated, it means that a packet has been duplicated, so the extra copies can be discarded. If there is a gap, it means that a packet is missing. It may come later, in which case it is inserted in its rightful place. Or it may not come, in which case the sender must retransmit it. The sequence number also plays a role in flow control, a process whereby the receiver can limit the amount of data sent to it.

Since TCP is a connection-oriented protocol it must acknowledge receipt of data. This is the purpose of the *acknowledgment number* field, which is used by the receiver to acknowledge the packets received from the sender. It has meaning only when the ACK (acknowledgment) bit in the control field (see next paragraph) is set, which is the usual case when TCP is transferring information.

Next is the *offset* field, which tells the receiver where in the packet the data starts. Since there can be options in the TCP header, the receiver must know where to start looking for data. The *control* field consists of 12 single-bit flags, of which the first four have no meaning at present. The remaining flags signal such things as congestion status, use of the acknowledgment and urgent fields, sequence number synchronization, and a flag to indicate that the sender is finished sending.

The next field, the *window*, indicates how much data the receiving system is capable of accepting. This is a critical flow-control mechanism. It is quite conceivable that a machine receiving data will not be able to process it as fast as the sender can send it. For example, suppose a file server at the headend is downloading audio files to a small device, such as an MP3 player. Over a fast

connection, it is quite possible that the server can get data to the MP3 player faster than it can store it in memory. The MP3 player must have some way to say, "Slow down." It decreases the size of its window when it acknowledges a packet, telling the sending machine to send less data next time. This takes care of end-to-end flow-control issues, but shortly we shall discuss another flow-control issue, that of congestion in the network.

The *checksum* detects errors in the TCP packet. The last part of the non-optional TCP header is the *urgent pointer*. It can be used to notify the receiving system that urgent data is coming and that the receiving system needs to act on that data as quickly as it can. The value in the field indicates the end of the urgent data. For example, suppose you are receiving something from a sender and determine that your machine is not reacting properly, so you need to stop the transmission. You send a "stop" command to the server sending the information, and the urgent flag is set to tell the server to act on that packet as quickly as it can.

The *TCP options* field, if used, can carry certain information that may or may not be needed. These options include a window scale factor, time stamps (for time-critical data), and a message digest of the packet. Application data makes up the rest of the data in the TCP packet.

When a host needs to send TCP data, it must first set up a session with the opposite end (recall that a characteristic of connection-oriented protocols is that a session, or *call*, must be set up before data can be passed and must be terminated, or *torn down*, at the end). The process of setting up the session is initiated by the host, which sends a *connection request* message to the intended receiver. The intended receiver then returns a *connection confirm* message, after which the sender returns an *acknowledgment* message. At the completion of this so-called *three-way handshake*, the connection is set up and data may be passed. Similarly, when the information has been passed, the connection must be closed in each direction.[20–22]

Congestion Control

Several times we have referred to congestion and control of congestion. Congestion occurs when something in the network gets data faster than it can process it. The receiving host can become congested if it gets data faster than it can handle it, and this situation is accommodated by the window field described earlier in the TCP section. If, on the other hand, a router gets congested, it can, if possible, set the ECN bits in the IP header, which will cause the receiver to slow down data delivery.

Two common techniques for congestion control at the TCP level are called *slow-start* and *multiplicative decrease*. Notice the transfer rate while downloading a large file, such as the radio frequency interface (RFI) files from *www.cablemodem.com*. You will see that the file transfer rate starts slowly, increases rapidly, possibly backing off, and eventually reaches a more-or-less steady state. You are watching these congestion mitigation techniques at work.

Slow-start avoids swamping the Internet with additional traffic immediately after congestion clears or when a new connection starts. When the file transfer

starts, the sender transfers data at a relatively slow rate, starting by transferring a single segment and waiting until that segment is acknowledged. When that first segment is acknowledged, the sender transmits two more segments. When they are acknowledged, it transmits two more, then four more, and so on. Eventually either the receiver's limit is reached or receipt of acknowledgments slows, indicating congestion somewhere along the route.

If congestion is encountered after the file transfer begins, multiplicative decrease is invoked. The number of segments transmitted is halved, and retransmission delayed, for every acknowledgment *not* received. This forces the sender to rapidly cut back to remove congestion, hopefully before anyone's packets are dropped. After the sender determines that no more packets are being dropped, it enters the slow-start sequence again, allowing the flow rate to increase again. In order to further avoid recreating congestion, as soon as the transmission rate reaches half the rate prior to congestion, the rate of increase is slowed.

5.4 Summary

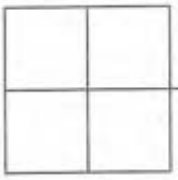
This chapter has attempted to introduce the key points related to data communications on hybrid fiber coaxial networks. It is not intended as a course in data communications, and many important topics in data communications were omitted. Those included were selected because of their particular relevance to data communications as it is being practiced on cable television systems today. One can expect many changes in techniques in the next few years as the technology matures.

The next chapter will deal with one particular system that transmits data on HFC plant, the cable telephone system. Many of the concepts presented in this chapter will be useful there.

Endnotes

1. Larry Peterson et al., *Computer Networks*. San Francisco: Morgan Kaufmann, 1996, p. 35ff.
2. Jim Stratigos, *ISO/OSI Model Introduction and Ethernet Overview*. Preconference tutorial presented at the 1997 Conference on Emerging Technologies, Nashville, TN, by the Society of Cable Telecommunications Engineers.
3. Jean Walrand and Pravin Varaiya, *High-Performance Communications Networks*. San Francisco: Morgan Kaufmann, 1996, p. 92ff.
4. An interesting historical perspective on the Internet, which explains the routing concept somewhat incidentally, may be found in Katie Hafner and Matthew Lyon, *Where Wizards Stay Up Late*. New York: Simon & Shuster, 1996.

5. http://www.cablemodem.com
6. Society of Cable Telecommunication Engineers, *Data-Over-Cable Systems Radio Frequency Interface Specification 1.1*, SCTE 23-1 2002.
7. http://www.usb.org
8. http://www.pcisig.com
9. See http://www.usb.org, *PCI-X 2.0 Overview*.
10. Cisco Systems, *Migrating Simple Data over Cable Services to DOCSIS 1.1*. Available at www.cisco.com.
11. Erich Arnold, *Content on Demand: How DOCSIS 1.1 Dynamic Service Flows Deliver the Goods*. Western Show, 2001.
12. Jean Walrand and Pravin Varaiya, op cit.
13. http://www.ieee.org
14. ANSI/TIA/EIA 570-A, *Residential Telecommunications Cabling Standard*. Available at http://global.ihs.com. See Figure 2 for an example of proper residential wiring. See also Endnote 15.
15. BICSI, *Residential Network Cabling*. 2002. BICSI is a not-for-profit telecommunications association supporting the design and distribution of telecommunications wiring in buildings. See www.bicsi.org. The book may also be obtained from commercial booksellers.
16. Charles E. Spurgeon, *Ethernet: The Definitive Guide*. Tampa: O'Reilly & Associates, 2000. Sebastopol, CA.
17. David Cunningham and William Lane, *Gigabit Ethernet Networking*. New York: Macmillan Technical, 1999.
18. Stephen A. Thomas, *IP Switching and Routing Essentials*. New York: Wiley, 2002.
19. http://www.ehsco.com/reading/19960515ncw3.html. See also www.webopedia.com.
20. Stephen A. Thomas, op. cit.
21. Douglas I. Comer, *Internetworking with TCP/IP*, 3rd ed. Upper Saddle River, N.J.: Prentice-Hall, 1995.
22. Floyd Wilder, *A Guide to the TCP/IP Protocol Suite*, 2nd ed. Boston: Artech House, 1998.



Chapter 7
Signal Reception

7.1 Introduction

This chapter is the first of three devoted to the headend. Chapter 7 covers the ways in which signals get into the headend. Chapter 8 will cover the electronics used in processing the signals once they get into the headend, and Chapter 9 will cover topics related to the signals after they have been processed.

Chapter 7 deals primarily with antennas of various types used to capture signals for the headend, and with propagation of signals from the source to those antennas.

Although cable operators usually use antennas to receive signals, antennas are sometimes described in terms of their operation when transmitting signals. The principle of duality applies, which states that an antenna is capable of operating as either a radiator or receptor of power. This principle is based on Maxwell's equations, which describe the co-existence of radiated fields and those contained within a conductor, without defining which causes the other.

7.2 Off-Air Reception

Cable television got its start as a community antenna service, in which a large antenna was erected in a favorable place and signals from it were supplied to the community. Today, off-air channels still tend to be some of the more watched though other channels are also popular. Some cities remain prime candidates for a community antenna service since propagation conditions don't lend themselves to good off-air reception at all locations. Cities located in mountainous terrain are good examples. As cable and broadcast television have learned to cooperate, it is becoming more common for cable headends to be supplied with a direct feed from the broadcaster. That way, the so-called perils of the radio path don't take their toll on the signals. Even so, it is still common for local and distant broadcasters to be picked up using off-air antennas.

7.2.1 Antennas for Reception of Off-Air Signals

Most, but not all, off-air antennas are of one of two types: Yagis or log periodic arrays. However, before we discuss them, it is necessary to discuss the dipole, which is fundamental to the operation of many other types of antennas. The dipole will be used to introduce certain concepts that are useful throughout the chapter.

The Dipole Antenna

A dipole antenna consists of a conductor broken in the middle to allow it to be fed from a balanced transmission line. The total length of the dipole is normally one-half wavelength though other lengths are used in certain cases.

Figure 7.1 illustrates operation of a basic half-wave resonant dipole antenna. Shown in Figure 7.1(a) is a dipole antenna excited from an RF source. The balanced nature of the ideal circuit ensures that identical current will exist on each side of the dipole. Current near the ends of the dipole is of necessity small,

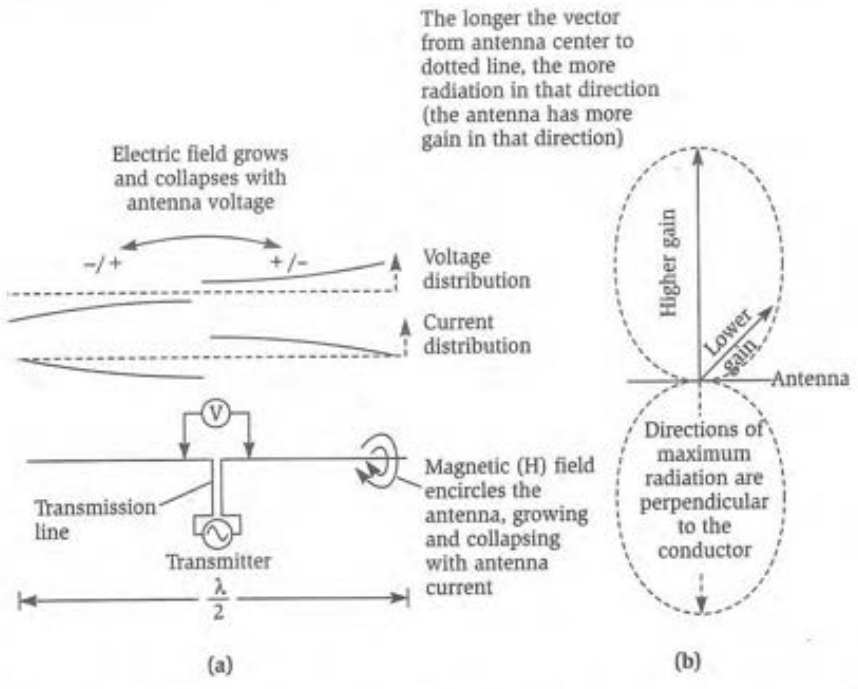


Figure 7.1 Operation of basic dipole antenna. (a) Half-wave dipole with current distribution and oscillating doublet. (b) E plane pattern.

vanishing at the ends, because electrons, whose movement constitutes the antenna current, have nowhere to move at the very end of the wire. As you move from either end toward the middle, the RMS current magnitude increases until it reaches a maximum at the center. Similarly, the RMS voltage measured from one side of the antenna to the like position on the other side (if it could be so measured) is lowest in the middle and rises to a maximum at the two ends.

Instantaneously, the dipole is charged negatively on one side, beginning at zero and rising to a maximum charge proportional to the power supplied; then the charge decreases to zero, and that side of the antenna becomes charged positively on the next half-cycle of the exiting waveform. This process creates a rising and falling electric field from one side of the dipole to the other, which moves away from the antenna. Similarly, the current in the dipole establishes a magnetic field encircling the dipole as shown, which also moves away from the antenna. The electric and magnetic fields together form the radiated electromagnetic field.

Maxwell's equations form the fundamental mathematics describing the action of antennas and the radiation of electromagnetic energy. They say that if an alternating electric field exists, then an alternating magnetic field must exist, and vice versa. The relative magnitude of each depends on the medium in which the electromagnetic field is propagating.

Polarization

Notice that the dipole, when parallel to the surface of the earth, radiates an electric field that is horizontal. Thus, it is said that the radiated field is *horizontally polarized*. The polarization ascribed to a radiated field is always that of the electric field. It is possible to position the dipole such that the electric field is vertically polarized. If two dipoles are placed at right angles and fed with the same signal, but 90° out of phase, then the electric field (along an axis perpendicular to the plane containing the two dipoles) will rotate between vertical and horizontal polarization, and the signal is said to be *circularly polarized*. Depending on which of the two dipoles is fed with the leading phase, the circularly polarized field will rotate either clockwise or counterclockwise (viewed with the field approaching the observer), resulting in a field that is said to be either *clockwise* or *counterclockwise* circularly polarized.

Near and Far Fields

Figure 7.2 illustrates a horizontally polarized wave front in the far field. Engineers often talk about *far field* and *near field* signals. Most computations assume far field conditions. Though there is no strict definition of far and near field, in order to be considered a far field signal, the electric and magnetic fields must be in a plane at right angles to the direction of travel of the wave, and the signal must be diverging according to the square law; that is, the signal power in a

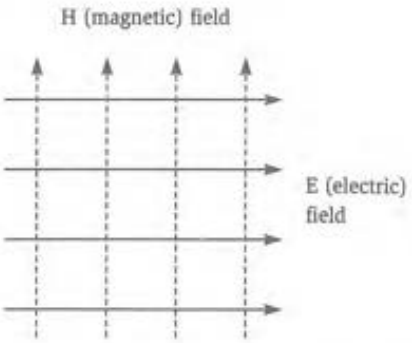


Figure 7.2 Plane wave traveling into page.

given cross-sectional area (the "power density") must be quartered every time the distance from the source is doubled. In general, these conditions are not met until the wave has traveled from the antenna by a few times the wavelength.

Note in Figure 7.1 that the E field close to the antenna is not quite flat (contained in a plane), and the H field is circling the antenna, so it cannot lie in a plane. As you go farther from the antenna, the diameter of the H field gets so large that it can be considered to lie in a plane for all practical purposes. Any magnetic field must close on itself, so the H field forms a circle. However, the dimensions of the circle become so large that for all practical purposes the H field lies in the same plane as that occupied by the E field.

Another concept of near and far field is that, in the near field, the electromagnetic field is influenced differently by different parts of the antenna, depending on the distance from the antenna part to the point of observation. That is, the antenna cannot be assumed to be at one point with respect to the point of observation. As you move far enough from the antenna, the antenna becomes more of a point source so that each part of the antenna influences the field in the same proportion as does any other part.

Dipole Pattern—E Plane

Figure 7.1(b) shows a top view of the horizontal half-wave dipole antenna, with the radiation pattern it produces. The pattern is much different if the dipole is any length other than one-half wavelength.[1] The dashed lines illustrate the relative radiation in each direction. In a direction in which the line is farther from the antenna, the antenna radiates more signal. As shown, the radiation from a half-wave dipole is strongest in a direction at right angles to the direction of the dipole. Note that Figure 7.1(b) is drawn looking down on a

horizontally polarized dipole, which has its E field parallel to the ground, and so is said to be horizontally polarized. Because the figure is drawn showing the radiation pattern such that the E field is in the plane of the pattern, this antenna pattern is said to be an *E field* pattern. Had the pattern been drawn looking at the end of the dipole, we would say that the pattern was the *H field* pattern.

Dipole Pattern—H Plane

In free space, a dipole would radiate equally in every direction when viewed end-on (the H field pattern). In the real world, dipoles are often operated close enough to ground that ground influences the pattern, and the radiation pattern would depend on how far above the ground plane the antenna was positioned. For example, a dipole located $\lambda/4$ (one-quarter wavelength) above a perfect ground has its maximum radiation at 90° (straight up). If the antenna is located at a height of $\lambda/2$, the maximum radiation is at about 30°. At a height of λ, the maximum vertical radiation angle is about 15°. Above real ground, the effect on the pattern is different, depending on ground conductivity. The vertical pattern of the antenna has to be taken into account when using a dipole to estimate the field strength of a signal.[2]

Gain of a Dipole

Since an antenna cannot amplify a signal, gain of an antenna can come only from its directionality. It must rob power from one direction of radiation in order to transmit it in another. The pattern shown in Figure 7.1 shows some directivity associated with a dipole, which means that *with respect to an isotropic radiator*, a dipole does have gain. (An isotropic radiator radiates equally in all directions.) A dipole can be shown to exhibit gain in the direction of maximum radiation, of 2.15 dBi,[3] where the suffix i indicates that the gain is stated with respect to an isotropic radiator. Antenna gain is sometimes stated with respect to isotropic, and sometimes with respect to a dipole, dBd. Thus, the gain of a dipole may be expressed as either 2.15 dBi or 0 dBd.

Practical Dipole Antennas

Figure 7.3 illustrates some practical aspects of dipole antennas. This is of interest to the cable operator because he or she is sometimes called upon to use dipole antennas to make rough field strength measurements. The length of a half-wave dipole is given by dividing the velocity of propagation *in the antenna wire*, by twice (to make the dipole one-half wavelength) the frequency of the signal to be received and/or transmitted. The velocity of propagation in the antenna wire is a function of several things, but at lower frequencies with thin wire antennas, the length may be approximated by

$$l \approx \frac{468}{f} \tag{7.1}$$

where

l = antenna length (feet)
f = frequency (MHz)

As the frequency increases, the length-to-diameter ratio drops, and the resonant length of the antenna tends to decrease compared with that given in the formula. Capacitive loading may also shorten the antenna. Since a resonant dipole in free space has an input impedance very close to 75 ohms, the return loss of the antenna can be measured with common cable equipment, and the length of the antenna adjusted until the return loss is optimized for the frequency at which the dipole is to be used. At lower frequencies, the dipole will have to be suspended from its ends, so end insulators are used to attach support cables. Depending on the support material, capacitive coupling across the end insulators may lower the resonant frequency somewhat.

In order to maintain the required balance (with respect to ground) between the two sides of the antenna, it must be fed from a balanced source (or must supply signal to a balanced load). Most equipment used in cable work is unbalanced, so it is necessary to convert from unbalanced to balanced using some sort of *balun* (*bal*ance to *un*balance). Several types of baluns are possible; the one shown is often useful at television frequencies. The coaxial cable closest

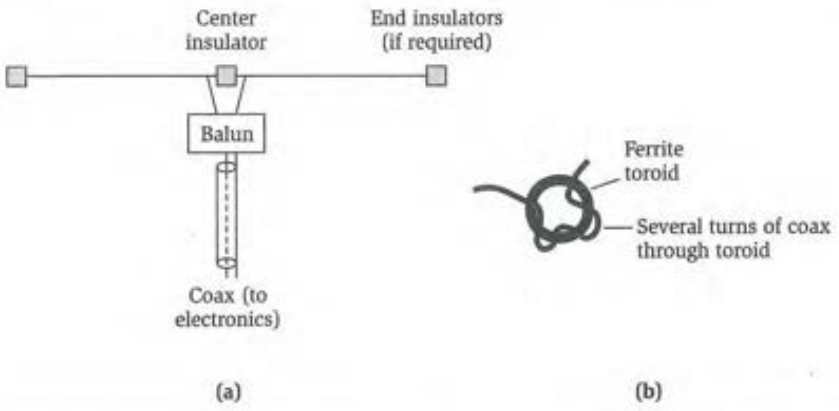


Figure 7.3 Practical dipole. (a) Practical dipole configuration. (b) One type of balun used to balance current.

to the antenna feedpoint is wrapped several times through a ferrite toroid. The resulting inductive reactance raises the impedance to any common mode current that may exist in the coax. (Common mode current is that which is in phase in the shield and center conductor of the coax.) Differential mode current, where the current in the center conductor and that in the shield are equal and out of phase, doesn't see the inductive reactance. Failure to ensure balance at the antenna driving point can cause distortion of the antenna pattern, and pickup on the coax, leading to errors in measured field strength.

A signal level meter and a half-wave dipole can be used to estimate the field strength of a radiated signal. The level of a signal is given by[4]

$$E \approx 21 f 10^{S/20} \tag{7.2}$$

where

E = field strength (μV/m)
f = frequency (MHz)
S = signal level (dBmV)

This relation is valid only for a resonant half-wave dipole in free space, oriented for maximum signal strength, and measured in a 75-ohm environment. Of course, cable and balun loss must be accounted for. Although mathematically accurate, this measurement is represented to be only an approximation to the true field strength because a number of variables can combine to reduce the accuracy. Accurate field strength measurements are difficult to make and require careful calibration of the equipment and the site. Unaccounted-for reflections alone can change the measurement by several decibels.

Even though a dipole may be used for reception of television signals, it is more common to employ a more directive antenna, which will exhibit additional gain. The gain is sometimes useful, and the directivity is almost always useful to reduce ghosting. Two types of antennas are used commonly in professional receive applications. Yagi antennas are used for single-channel reception, or at higher frequencies a Yagi may cover several channels. At lower frequencies, it is not possible to build a Yagi with enough bandwidth to cover several channels, but log periodic antennas may be used. Occasionally, other antennas will be used, but they are less common.

Yagi Antennas

The Yagi (or Yagi–Uda) antenna is named after its inventors. It is a parasitic antenna, meaning that one of several elements is driven, and current in the driven element produces current in the others. Figure 7.4 illustrates a four-element Yagi. Yagis have at least two elements, and the number can go much