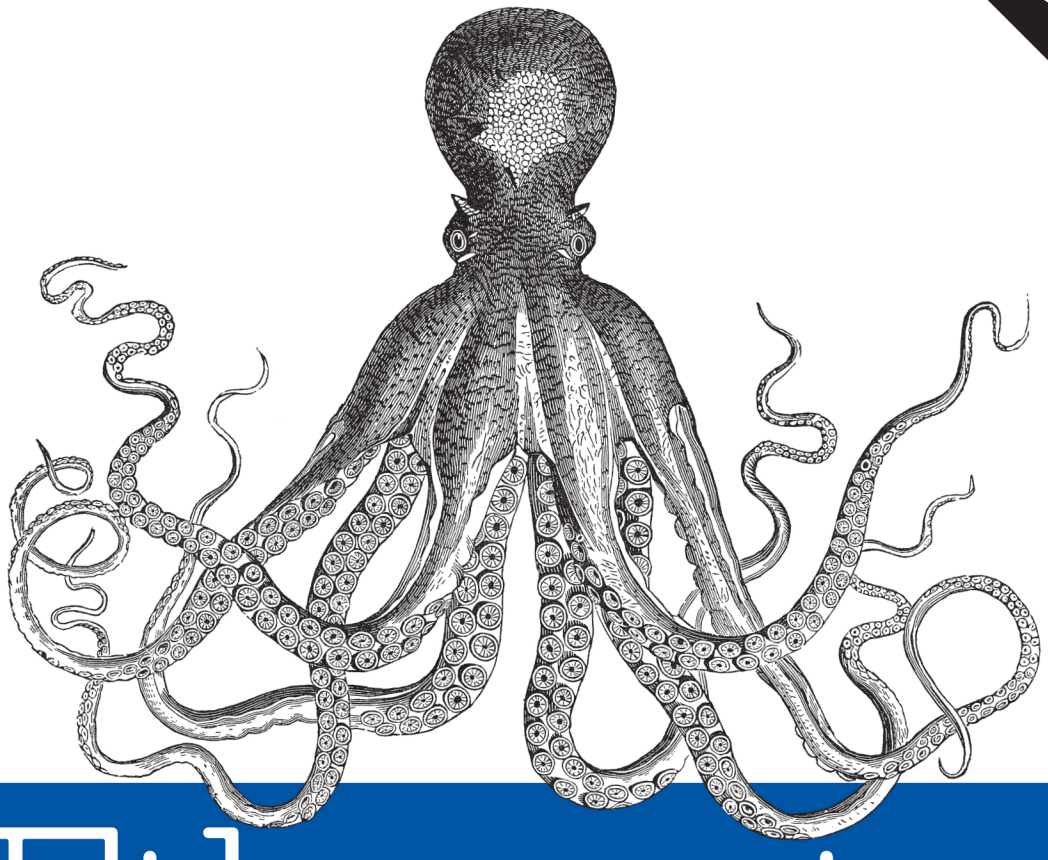
O'REILLY®

2nd Edition



Ethernet

The Definitive Guide

DESIGNING AND MANAGING LOCAL AREA NETWORKS

Charles E. Spurgeon &
Joann Zimmerman

O'REILLY®

Ethernet: The Definitive Guide

Get up to speed on the latest Ethernet capabilities for building and maintaining networks for everything from homes and offices to data centers and server machine rooms. This thoroughly revised, comprehensive guide covers a wide range of Ethernet technologies, from basic operation to network management, based on the authors' many years of field experience.

When should you upgrade to higher speed Ethernet? How do you use switches to build larger networks? How do you troubleshoot the system? This book provides the answers. If you're looking to build a scalable network with Ethernet to satisfy greater bandwidth and market requirements, this book is indeed the definitive guide.

- Examine today's most widely used media systems and advanced 40 and 100 gigabit Ethernet
- Learn about Ethernet's four basic elements
- Explore full-duplex Ethernet, Power over Ethernet, and Energy Efficient Ethernet
- Understand structured cabling systems and the components you need to build your Ethernet system
- Use Ethernet switches to expand and improve network design
- Delve into Ethernet performance, from specific channels to the entire network
- Get troubleshooting techniques for problems common to twisted-pair and fiber optic systems

"Heck, I designed some of the technologies described in this book, but I still keep a copy of *The Definitive Guide* nearby for reference!"

—Rich Seifert
Long-time Ethernet developer and author of *The Switch Book* (Wiley) and *Gigabit Ethernet* (Addison-Wesley)

Charles Spurgeon, a senior technology architect at the University of Texas at Austin, works on the network system serving over 70,000 users in 200 buildings. He helped build prototype Ethernet routers that became the founding technology for Cisco Systems.

Joann Zimmerman, a former software engineer, has written and documented compilers, software tools, and network monitoring software. She created the build and configuration management process for several companies.

NETWORKING/NETWORK ADMINISTRATION

US $44.99 CAN $47.99

ISBN: 978-1-449-36184-6

Twitter: @oreillymedia
facebook.com/oreilly

SECOND EDITION

Ethernet: The Definitive Guide

Charles E. Spurgeon and Joann Zimmerman

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

Ethernet: The Definitive Guide, Second Edition

by Charles E. Spurgeon and Joann Zimmerman

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Meghan Blanchette
Production Editor: Nicole Shelby
Copyeditor: Rachel Head
Proofreader: Jasmine Kwityn

Indexer: Judy McConville
Cover Designer: Randy Comer
Interior Designer: David Futato
Illustrator: Rebecca Demarest

March 2014: Second Edition

Revision History for the Second Edition:

2014-03-11: First release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449361846* for release details.

ISBN: 978-1-449-36184-6

[LSI]

Table of Contents

Part V. Performance and Troubleshooting

Preface

This is a book about Ethernet, the world's most popular network technology, which allows you to connect a variety of computers together with a low-cost and extremely flexible network system. Ethernet is found on a wide variety of devices, and this widespread support, coupled with its low cost and high flexibility, are major reasons for its popularity.

The Ethernet standard has grown to over 3,700 pages, and it covers a multitude of Ethernet technologies designed for multiple environments. Ethernet is used to build home networks, office and campus network systems, as well as wide area networks that span cities and countries. There are Ethernet systems designed for networking a neighborhood, as well as Ethernets designed for networking inside automobiles to link the multiple computers found there these days.

The goal of this book is to provide a comprehensive and practical source for information on the most widely used Ethernet technologies in a single volume. This book describes the varieties of Ethernet commonly used in homes, offices, and campus networks, as well as several systems typically used in data centers and server machine rooms. These include the most widely used set of Ethernet media systems: 10 Mb/s Ethernet, 100 Mb/s Fast Ethernet, and 1000 Mb/s Gigabit Ethernet, as well as 10 Gigabit and 40 and 100 Gigabit Ethernet. We also describe full-duplex Ethernet, Ethernet Auto-Negotiation, Power over Ethernet, Energy Efficient Ethernet, structured cabling systems, network design with Ethernet switches, network management, network troubleshooting techniques, and more.

To provide the most accurate information possible, we referred to the complete set of official Ethernet standards while writing this book. Our experience includes working with Ethernet technology since the early 1980s, and many hard-won lessons in network design and operation based on that experience have made their way into this edition.

Ethernet Is Everywhere

Ethernet is the most widely used networking technology, and Ethernet networks are everywhere. There are a number of factors that have helped Ethernet to become so popular. Among these factors are cost, scalability, reliability, and widely available management tools.

Cost

The rapid evolution of new capabilities in Ethernet has been accompanied by an equally rapid decrease in the cost of Ethernet equipment. The widespread adoption of Ethernet technology created a large and fiercely competitive Ethernet marketplace, which serves to drive down the cost of networking components. The consumer wins out in the process, with the marketplace providing a wide range of competitively priced Ethernet components to choose from.

Scalability

The first industry-wide Ethernet standard was published over 30 years ago, in 1980. This standard defined a 10 megabits per second (Mb/s) system, which was very fast for the time. The development of the 100 Mb/s Fast Ethernet system in 1995 provided a tenfold increase in speed. Following on that success came the development of twisted-pair Gigabit Ethernet in 1999. Network interfaces that can automatically support 10, 100, and 1000 Mb/s operation of twisted-pair media systems are widely available, making the support of high-performance networking easy to accomplish.

Applications tend to grow to fill all available bandwidth. To manage the constant increase in network usage, the 10 Gigabit Ethernet standard was developed in 2002, and most recently the 40 and 100 Gigabit systems were standardized in 2010. All of this progress in Ethernet capabilities makes it possible for a network manager to provide high-speed backbone systems and connections to high-performance servers.

Desktop machines can be connected to an Ethernet link that can operate at 10 Mb/s Ethernet, 100 Mb/s Fast Ethernet, or Gigabit Ethernet speeds, as required. Network routers and switches can use 10 Gigabit and 40 or 100 Gigabit links for network backbones, and data centers can connect to high-performance servers at 10, 40, or even 100 gigabits per second (Gb/s).

Reliability

Ethernet is simple and robust and reliably delivers data day in and day out at sites all over the world. Ethernet based on twisted-pair media was introduced in 1987, making it possible to provide Ethernet signals over a structured cabling system.

Structured cabling provides a data delivery system for a building that is modeled on high-reliability cabling practices originally developed for the telephone system. This makes it possible to install a standards-based cabling system for Ethernet that is highly reliable and easy to manage.

Widely Available Management Tools

The widespread acceptance of Ethernet brings with it the wide availability of Ethernet management and troubleshooting tools. Management tools based on standards such as the Simple Network Management Protocol (SNMP) make it possible for network administrators to keep track of an entire campus full of Ethernet equipment from a central location. Management capabilities embedded in Ethernet switches and computer interfaces provide powerful network monitoring and troubleshooting capabilities.

Design for Reliability

A major goal of this book is to help you design and implement reliable networks, because network reliability is of paramount importance to users and organizations. Access to the Internet and information sharing between networked computers is an essential feature of today's world, and if the network fails, everything comes to a halt. This book shows you how to design reliable networks, how to monitor them and keep them working reliably, and how to fix them should something fail.

The wide range of Ethernet components and cabling systems available today provides enormous flexibility, making it possible to build an Ethernet to fit just about any circumstance. However, all this flexibility does have a price. The many varieties of Ethernet each have their own components and their own configuration rules, which can make the life of a network designer complex. Designing and implementing a reliable Ethernet system requires that you understand how all the bits and pieces fit together, and that you follow the official guidelines for the configuration of the media systems. To help you with that task, this book provides the configuration guidelines for the widely used media systems.

Downtime is Expensive

Avoiding network downtime is important for a number of reasons, not least of which is the cost of a network outage. Some quick "back of the envelope" calculations can show how expensive network downtime can be. Let's assume that there are 1,000 network users at the Amalgamated Widget Company, and that their average annual salary including all overhead (benefits, etc.) is $100,000. That comes to $100 million a year in employee costs.

Let's further assume that everyone in the company depends on the network to get their work done, and that the network is used 40 hours a week, for about 50 weeks of the year.

That's 2,000 hours of network operation. Dividing the annual employee cost by the hours of network operation shows that the network is supporting $50,000 per hour of employee cost during the year.

Let's further assume that when we total up all of the network outages over the period of a year in our hypothetical corporation, we find that the network was down just 1% of the time (99% uptime, or "two nines"). That sounds like really good uptime, but that small fraction of 2,000 hours represents a total of 20 hours of network outage. Twenty hours of network downtime at $50,000/hour is $1,000,000 in lost productivity due to network outage.

Obviously, our example is very "quick and dirty." We didn't bother to calculate the impact of network outages during times when no one is around but when the network is still nevertheless supporting critically important servers. Also, we're assuming that a network failure brings all operations to a halt, instead of trying to factor in the varying effects of localized failures that cause outages on only a portion of the network system. Nor do we try to estimate how much other work people could get done while the network is down, which would tend to lessen the impact.

However, the main point is clear: even relatively small amounts of network downtime can cost quite a lot in lost productivity. That's why it's worth investing extra time, effort, and money to create the most reliable network system you can afford.

How to Use This Book

The goal of this book is to provide the information needed for you to understand and operate any Ethernet system. For example, if you are a newcomer to Ethernet and you need to know how twisted-pair Ethernet systems work, then you can start with Part I. After reading those chapters, you can go to the twisted-pair media chapters in Part II, as well as the twisted-pair cabling information in Part III. Twisted-pair cables are connected together to form a network using switches, and these are described in Part IV.

Experts in Ethernet can use the book as a reference guide and jump directly to those chapters that contain the information they need.

Organization of This Book

The purpose of this book is to provide a comprehensive and practical guide to the Ethernet system and the Ethernet devices and components commonly used in office and building networks. The emphasis is on practical issues, with minimal theory and jargon. Chapters are kept as self-contained as possible, and many examples and illustrations are provided. The book is organized into six parts to make it easier to find the specific information you need.

Here's what you'll find in each of these parts:

- Part I provides an introduction to the Ethernet standard and a description of Ethernet theory and operation. The chapters in this part cover those portions of Ethernet operation that are common to all Ethernet media systems, including the Ethernet frame, the operation of the media access control system, full-duplex mode, and the Auto-Negotiation protocol.
- Part II contains a description of each of the Ethernet media systems. It begins with the basics of Ethernet media signaling in Chapter 7, which also covers the Energy Efficient Ethernet system that saves power by modifying the media signaling during idle periods. Chapters 8 through 14 describe specific media systems, including 10, 100, and 1000 Mb/s, and 10, 40, and 100 Gb/s systems.
- Part III offers a description of structured cabling systems and the components and cables used in building your Ethernet system, including a discussion of the structured cabling standards and details on twisted-pair and fiber optic cabling.
- Part IV describes the fundamentals of network design, including how to design and build Ethernet systems using Ethernet switches.
- Part V covers Ethernet performance and troubleshooting.
- Part VI contains the appendixes and glossary.

Disclaimer

While every precaution has been taken in the preparation of this work, the authors assume no responsibility for errors or omissions, or for damages resulting from the use of information contained herein. We make no claims about the completeness or the accuracy of the information as it may apply to any field conditions.

Conventions Used in This Book

Italic
: Used for filenames, new terms, and URLs.

This icon designates a note, which is an important aside to its nearby text.

This icon designates a warning relating to the nearby text.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *http://oreil.ly/ethernetTDG_2e*.

To comment or ask technical questions about this book, send email to *bookquestions@oreilly.com*.

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Acknowledgments

This book would not have been possible without the help of many people. First and foremost, the authors would like to thank the inventors of Ethernet, Bob Metcalfe and his fellow researchers at Xerox PARC. Their work revolutionized the way computers are used, unleashing a powerful new communications technology based on information sharing on computers linked with networks. We also thank the many engineers who have voluntarily given their time in countless IEEE standards meetings to develop new capabilities for the Ethernet system and to write the Ethernet specifications.

The authors would also like to thank our acquisitions editor at O'Reilly, Meghan Blanchette, and the other editors and staff of O'Reilly who have worked on this book, for their assistance and attention to detail. We'd also like to thank Tim O'Reilly for creating a technical publishing house that supports such a wide variety of information resources, and that treats both readers and writers with respect.

Finally, we'd like to thank Rich Seifert, author of *The Switch Book*, engineer and developer of Ethernet technology, and a participant in the creation of Ethernet standards from the earliest days of Ethernet. Rich provided in-depth reviews of the manuscript that are very much appreciated and that helped improve the final work. Of course, the authors alone are responsible for any errors.

PART I
Introduction to Ethernet

The first part of this book provides a tour of basic Ethernet theory and operation. These chapters cover the portions of Ethernet operation that are common to all Ethernet media systems, including the Ethernet frame, the operation of the media access control system, full-duplex mode, and the Auto-Negotiation protocol.

CHAPTER 1

The Evolution of Ethernet

Ethernet is used to build networks from the smallest to the largest, and from the simplest to the most complex: it connects home computers and other household devices, but it also connects the building networks that support servers and wired desktop computers, as well as the wireless access points that support smartphones, laptops, and tablets. Ethernet provides the connections that make up the worldwide Internet and that connect the Internet to our workplaces and our homes.

Ethernet's longevity is remarkable. The memo describing the network technology that became Ethernet was written in May 1973. There have been many changes as computers have evolved over the years, but Ethernet continues to be the network technology of choice. This is because Ethernet has been constantly reinvented, evolving new capabilities to stay current with the rapid transformations in the computer industry and, in the process, becoming the most widely used network technology in the world.

History of Ethernet

On May 22, 1973, while working at the Xerox Palo Alto Research Center (PARC) in California, Bob Metcalfe wrote a memo describing the network system he had invented for interconnecting advanced computer workstations called Xerox Altos, making it possible to send data between them and to high-speed laser printers. The Xerox Alto was the first personal computer workstation with graphical user interfaces and a mouse pointing device. The PARC inventions also included the first laser printers for personal computers and, with the creation of Ethernet, the first high-speed *local area network* (LAN) technology to link everything together.

This was a remarkable computing environment for the time, since the early 1970s was an era in which computing was dominated by large and expensive mainframe computers. Few places could afford to buy and support mainframes, and few people knew how

to use them. The inventions at Xerox PARC helped bring about a revolutionary change in the world of computing.

A major driver of this revolutionary change was the use of Ethernet LANs to enable communication among computers. Combined with the development of the Internet and the Web, this new model of interaction between computers brought a new world of communications technology into existence.

The Aloha Network

Bob Metcalfe's 1973 Ethernet memo describes a networking system inspired by an earlier experiment in networking called the Aloha network. The Aloha network began at the University of Hawaii in the late 1960s, when Norman Abramson and his colleagues developed a radio network for communication among the Hawaiian Islands. This system was an early experiment in the development of mechanisms for sharing a common communications channel—in this case, a common radio channel.

The Aloha protocol was very simple: an Aloha station could send whenever it liked, and then wait for an acknowledgment. If an acknowledgment wasn't received within a short amount of time, the station would assume that another station had transmitted simultaneously, causing a *collision* in which the combined transmissions were garbled so that the receiving station did not hear them and did not return an acknowledgment. Upon detecting a collision, both transmitting stations would choose a random backoff time, and then retransmit their packets with a good probability of success. However, as traffic increased on the Aloha channel, the collision rate would rapidly increase as well.

Abramson calculated that this system, known as *pure Aloha,* could achieve a maximum channel utilization of about 18%, due to the rapidly increasing rate of collisions under increasing load. Another system, called *slotted Aloha,* was developed that assigned transmission slots and used a master clock to synchronize transmissions; this increased the maximum utilization of the channel to about 37%. In 2007, Abramson received the IEEE's Alexander Graham Bell Medal for "contributions to the development of modern data networks through fundamental work in random multiple access."[1]

The Invention of Ethernet

Metcalfe realized that he could improve on the Aloha system of arbitrating access to a shared communications channel. He developed a new system that included a mecha-

1. The IEEE Global History Network biography of Norman Abramson states: "While at the University of Hawaii, he led efforts that gave rise to the construction and operation of the ALOHAnet, the first wireless packet network, and to the development of the theory of random access ALOHA channels. ALOHA channels have yielded significant advancements within wireless and local area networking, with versions still in use today in all major mobile telephone and wireless data standards. This influential work also developed the core concepts found today in Ethernet."

nism that detected when a collision occurred (*collision detection*). The system also included "listen before talk," in which stations listened for activity (*carrier sense*) before transmitting, and supported access to a shared channel by multiple stations (*multiple access*). Put all these components together, and you can see why the original channel access protocol specified for Ethernet is called Carrier Sense Multiple Access with Collision Detection (CSMA/CD). Metcalfe also developed a more sophisticated backoff algorithm, which, in combination with the CSMA/CD protocol, allowed the Ethernet system to function at up to 100% load.

In late 1972, Metcalfe and his Xerox PARC colleagues developed the first experimental "Ethernet" network system to interconnect Xerox Altos to one another, and to servers and laser printers. The signal clock for the experimental interface was derived from the Alto's system clock, resulting in a data transmission rate on the experimental Ethernet of 2.94 Mb/s.

Metcalfe's first experimental network was called the *Alto Aloha Network*. In 1973, Metcalfe changed the name to "Ethernet," to make it clear that the system could support any computer, not just Altos, and to point out that his new network mechanisms had evolved well beyond the Aloha system. He chose to base the name on the word "ether" as a way of describing an essential feature of the system: the physical medium (i.e., a cable) carries bits to all stations, much the same way that the old "luminiferous ether" was once thought to propagate electromagnetic waves through space.[2] Thus, *Ethernet* was born.

In 1976, Metcalfe drew the diagram shown in Figure 1-1, and it was used in his presentation at the National Computer Conference in June of that year. The drawing uses the original terms for describing Ethernet components.[3]

2. Physicists Albert Michelson and Edward Morley disproved the existence of the ether in 1887, but Metcalfe decided that it was a good name for his new network system that carried signals to all computers.

3. From *The Ethernet Sourcebook,* ed. Robyn E. Shotwell (New York: North-Holland, 1985), title page. Diagram reproduced with permission.

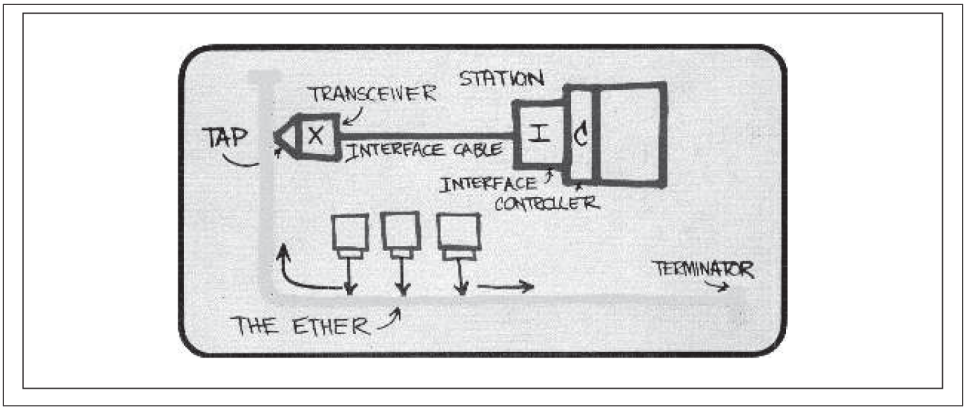


Figure 1-1. Drawing of the original Ethernet system

In July 1976, Bob Metcalfe and David Boggs published their landmark paper "Ethernet: Distributed Packet Switching for Local Computer Networks."[4] In late 1977, Robert M. Metcalfe, David R. Boggs, Charles P. Thacker, and Butler W. Lampson received U.S. patent number 4,063,220 on Ethernet for a "Multipoint Data Communication System with Collision Detection."

At this point, Xerox wholly owned the Ethernet system. The next stage in the evolution of the world's most popular computer network was to liberate Ethernet from the confines of a single corporation and make it a worldwide standard.

Reinventing Ethernet

No matter how well designed a network system is, it won't help you much if you can only use it with a single vendor's equipment. A network technology has to be supported by the widest variety of equipment possible to provide you with the greatest flexibility. For maximum utility, your network must be vendor-neutral (i.e., capable of interworking with all types of computers and other devices without being vendor-specific). This was not the way things worked in the 1970s, when computers were expensive and networking technology was exotic and proprietary.

Bob Metcalfe understood that a revolution in computer communications required a networking technology that everyone could use. In 1979, he set out to make Ethernet an open standard, and Xerox agreed to join a multivendor consortium for the purposes of standardizing an Ethernet system that any company could use. The era of open computer communications based on Ethernet technology formally began in 1980 when the Digital Equipment Corporation (DEC), Intel, and Xerox (DIX) consortium announced

4. *Communications of the ACM*, 19:7 (July 1976): 395–404.

the first standard for 10 Mb/s Ethernet. The original DIX standard was not copyrighted, allowing anyone to copy and use it.

This standard made the technology available to anyone who wanted to use it, producing an open system. As part of this effort, Xerox agreed to license its patented Ethernet technology for a mere $1,000 to anyone who wanted it. In 1982, Xerox also gave up its trademark on the Ethernet name. As a result, the Ethernet standard became the world's first open, multivendor LAN standard.

The idea of sharing proprietary computer technology in order to arrive at a common standard to benefit everyone was a radical notion for the computer industry in the late 1970s. It's a tribute to Bob Metcalfe's vision that he realized the importance of making Ethernet an open standard. As Metcalfe put it: "The invention of Ethernet as an open, non-proprietary, industry-standard local network was perhaps even more significant than the invention of Ethernet technology itself."[5]

In 1979, Metcalfe started a company to help commercialize Ethernet. He believed that computers from multiple vendors ought to be able to communicate compatibly over a common networking technology, making them more useful and, in turn, opening up a vast new set of capabilities for the users. *Com*puter *com*munication *com*patibility was the goal, leading Metcalfe to name his new company 3Com.

Reinventing Ethernet for Twisted-Pair Media

Ethernet prospered during the 1980s, but as the number of computers being networked continued to grow, the problems inherent in the original coaxial cable media system became more acute. Installing coaxial cables in buildings was a difficult task, and connecting computers to the cables was also a challenge.

A thin coaxial cable system was introduced in the mid-1980s that made it a little easier to build a media system and connect computers to it, but it was still difficult to manage Ethernet systems based on coaxial cable. Coaxial Ethernet systems employ a bus topology, in which every computer sends Ethernet signals over a single bus cable; a failure anywhere on the cable brings the entire network system to a halt, and troubleshooting a cable problem can take a long time.

The invention of *twisted-pair* Ethernet in the late 1980s, initially developed as a vendor innovation, made it possible to build Ethernet systems based on the much more reliable star-wired cabling topology, in which the computers are all connected to a central point.[6] These systems are much easier to install and manage, and troubleshooting is much easier and quicker as well. The use of twisted-pair cabling was a major change,

5. Shotwell, *The Ethernet Sourcebook*, p. xi.

6. The vendor was SynOptics Communications, whose LattisNet was the first twisted-pair product.

or reinvention, of Ethernet. Twisted-pair Ethernet led to a vast expansion in the use of Ethernet; the Ethernet market took off and has never looked back.

In the early 1990s, a structured cabling system standard for twisted-pair cabling systems in buildings was developed that made it possible to provide building-wide twisted-pair systems based on high-reliability, low-cost cabling adopted from the telephone industry. Ethernet based on twisted-pair media installed according to the structured cabling standard became the most widely used network technology. These Ethernet systems are reliable, are easy to install and manage, and support rapid troubleshooting for problem resolution.

Reinventing Ethernet for 100 Mb/s

The original Ethernet standard of 1980 described a system that operated at 10 Mb/s. This was quite fast for the time, but Ethernet interfaces in the early 1980s were expensive, due to the buffer memory and high-speed components required. Throughout the 1980s, Ethernet was considerably faster than the computers connected to it, making a good match between the network and the computers it supported. However, computer technology continued to evolve, and by the early 1990s ordinary computers had become fast enough to provide a major traffic load to a 10 Mb/s Ethernet channel.

Much to the surprise of those who thought that the original CSMA/CD-based Ethernet system was limited to 10 Mb/s, Ethernet was reinvented to increase its speed by a factor of 10. Based on technology developed by Grand Junction Networks (later acquired by Cisco Systems), the new standard created the 100 Mb/s Fast Ethernet system, which was formally adopted in 1995. Fast Ethernet provides both twisted-pair and fiber optic media systems, and it became widely adopted, first for network backbones and later for general computing.

With the invention of Fast Ethernet, multispeed twisted-pair Ethernet interfaces could be built, operating at either 10 or 100 Mb/s. These interfaces are able, through an Auto-Negotiation protocol, to automatically set their speed. This made the migration from 10 Mb/s to 100 Mb/s Ethernet systems easy to accomplish.

Reinventing Ethernet for 1000 Mb/s

In 1998, Ethernet was reinvented again, this time to increase its speed by another factor of 10. The Gigabit Ethernet standard describes a system that operates at the speed of 1 billion bits per second over fiber optic and twisted-pair media. The invention of Gigabit Ethernet made it possible to provide faster backbone networks as well as connections to high-performance servers.

The twisted-pair standard for Gigabit Ethernet provides high-speed connections to the desktop when needed. Multispeed twisted-pair Ethernet interfaces were built to operate

at 10, 100, or 1000 Mb/s, using the Auto-Negotiation protocol to automatically configure their speed.

Reinventing Ethernet for 10, 40, and 100 Gb/s

Not content to rest on its laurels, Ethernet has continued to expand beyond the original design constraints. Although it's not possible to support the original CSMA/CD shared-channel mode of operation at these higher speeds, that doesn't matter: virtually all Ethernet connections now operate in full-duplex mode, which does not rely on the CSMA/CD access control system.

The 10 Gb/s Ethernet standard, published in 2003, defined a set of fiber optic media systems operating at 10 billion bits per second. A twisted-pair 10 Gb/s standard was developed and published in 2006, providing 10 billion bits per second over Category 6A twisted-pair cables. Multispeed twisted-pair Ethernet interfaces can now operate at 10, 100, and 1000 Mb/s, and 10 Gb/s.

The 40 and 100 Gb/s Ethernet standard, which was published in 2010, defined both 40 and 100 Gb/s media systems. Since then, media systems have been evolving to carry 40 and 100 Gb/s Ethernet signals over fiber optic cables and short-range copper coaxial cables.

Reinventing Ethernet for New Capabilities

Ethernet innovations include not only new speeds and new media systems, but also new Ethernet capabilities. For example, the standardization of full-duplex Ethernet in 1997 made it possible for two devices connected over a full-duplex link to simultaneously send and receive data, thus allowing a 10 Gb/s link to provide a maximum of 20 Gb/s of data throughput.

The Auto-Negotiation standard complements the invention of twisted-pair Ethernet by providing the ability for switch ports and the computers connected to those ports to discover whether they support full-duplex mode and, if they do, to automatically select that mode of operation as well as automatically setting the highest link speed supported by both devices.

Another innovation has been the Power over Ethernet (PoE) standard, which uses the Ethernet cable that is providing data to also power the device connected to an Ethernet switch. This has become a widely adopted method for deploying wireless access points connected to Ethernet switch ports and drawing their power from the same cable that they use to send and receive Ethernet frames.

Ethernet Switches

The invention of full-duplex twisted-pair and fiber optic Ethernet coincided with the development of network switches, allowing network managers to build large networks based on switches and full-duplex links. Switches have Ethernet interfaces (ports), but the operation of switch protocols is not part of the Ethernet standard. Instead, the operation of switches is specified in the IEEE 802.1 series of standards, with the 802.1D standard providing the specifications for basic switches.

You can build a wide variety of networks with switches. There are switches designed for campus and enterprise networks, switches with special capabilities for data centers, switches that support carrier and long distance networks, and more.

Network design based on switches is a big topic with its own literature, based on the type of network being developed. There are books on campus and enterprise network design, as well as books on data center networks. This is a book on Ethernet standards and technology, and we don't have the space to provide an in-depth treatment of the 802.1 switch standards and the topic of network design with switches for multiple network types. However, Part IV, including Chapters 18 and 19, provides an introduction to switch operation and a discussion of how switches can be used in network designs.

The Future of Ethernet

Ethernet has come a long way since 10 Mb/s Ethernet became the world's first open standard for computer networking in the early 1980s. As you can see, the Ethernet system has been reinvented to provide more flexible and reliable cabling, to accommodate the rapid increase in network traffic with higher speeds, and to provide more capabilities for today's more complex network systems.

Ethernet has been able to meet these challenges while maintaining the same basic structure and operation, and doing it all at a reasonable cost. This fundamental stability, combined with the ability to evolve to meet new needs, is at the core of Ethernet's success.

CHAPTER 2

IEEE Ethernet Standards

Ethernet is standardized by the Institute for Electrical and Electronics Engineers (IEEE). The IEEE (pronounced "Eye-triple-E") is headquartered in New York City and has more than 425,000 members in over 160 countries. One of the largest worldwide professional organizations, the IEEE organizes conferences and publishes more than 150 transactions, journals, and magazines annually. The IEEE also develops standards in a broad range of industries, including telecommunications, information technology, nanotechnology, and power generation products and services.

The Ethernet standards produced by the IEEE Standards Association (IEEE-SA) are just one group of the more than 1,400 standards and projects under development. The IEEE-SA is composed of volunteers from the community of IEEE engineers and is not a formal part of any government. However, the IEEE standards are formally recognized by national standards groups (e.g., American ANSI, German DIN) and international standards organizations (e.g., ISO, IEC).

The process of developing IEEE standards involves engineers from industry, government, and other domains who volunteer their time to work together within the IEEE-SA framework to produce standards. In order to develop a set of specifications that participants agree will provide an open and interoperable standard that all vendors can use, the engineers are required to reach a consensus on the technical issues. The IEEE standards ensure that vendors can build equipment that works well together, thus expanding the marketplace and benefitting both manufacturers and consumers.

Evolution of the Ethernet Standard

The original 10 Mb/s Ethernet standard was first published in 1980 by the DEC-Intel-Xerox vendor consortium. Using the first initial of each company's name, this became known as the DIX Ethernet standard. This standard, entitled "The Ethernet, A Local Area Network: Data Link Layer and Physical Layer Specifications," contained the spec-

ifications for the operation of Ethernet as well as the specs for a single media system based on thick coaxial cable. As is true for most standards, the DIX standard was revised to add technical changes, corrections, and minor improvements. The last revision of this standard was DIX V2.0, published in November 1982.

At roughly the same time that the DIX standard was published, a new effort led by the IEEE to develop open network standards was also getting underway. Consequently, the original Ethernet technology, based on the use of a thick coaxial cable to provide a shared communications channel, ended up being standardized twice—first by the DIX consortium and a second time by the IEEE.

The IEEE standard is currently maintained by the IEEE 802 LAN/MAN Standards Committee (LMSC). According to the 2012 IEEE 802 LMSC Overview & Guide:

> *The first meeting of the IEEE, 'Local Area Network Standards Committee,' Project 802, was held in February of 1980. (The project number, #802, was simply the next number in the sequence being issued by the IEEE for standards projects.) There was originally only going to be one LAN standard, with speeds ranging from 1 to 20 Mb/s. It was later divided into a Media or Physical layer (PHY) standard, a Media Access Control (MAC) standard, and a Higher Level Interface (HILI) standard. The original access method was similar to that for Ethernet and used a passive bus topology.*

The IEEE 802.3 committee took up the network system described in the DIX standard and used it as the basis for the IEEE standard. The IEEE standard for Ethernet technology, "IEEE 802.3 Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications," was first published in 1985. Even though Xerox relinquished its trademark on the Ethernet name, the IEEE standard did not originally use "Ethernet" in the title. That's because the open standards committees were sensitive about using commercial names that might imply endorsement of a particular company. As a result, the IEEE called this technology *802.3 CSMA/CD*, or just *802.3*.[1] However, today the standard has dropped the use of "CMSA/CD" in the title, which has been changed to "IEEE Standard for Ethernet."

The IEEE 802.3 standard is the official Ethernet standard. From time to time, you may hear of other Ethernet "standards" developed by various groups or vendor consortiums. Or you may hear of a different technology, such as 802.11 wireless LANs, referred to as "Ethernet." However, if the technology isn't specified within the IEEE 802.3 standard, then it isn't officially Ethernet. That doesn't mean that the technology won't work, but it will typically be vendor-specific and not widely available from multiple vendors. It may also be a niche technology that was not considered useful enough to warrant inclusion in the standard.

1. Pronounced "eight oh two dot three."

The title of the most recent version of the IEEE standard as of this writing is: "IEEE Standard for Ethernet," IEEE Std 802.3-2012 (Revision of IEEE Std 802.3-2008)." The 2012 edition of the standard contains 3,747 pages and can be downloaded for free from the IEEE.

The abstract of the Ethernet standard reads:

> *Ethernet local area network operation is specified for selected speeds of operation from 1 Mb/s to 100 Gb/s using a common media access control (MAC) specification and management information base (MIB). The Carrier Sense Multiple Access with Collision Detection (CSMA/CD) MAC protocol specifies shared medium (half-duplex) operation, as well as full duplex operation. Speed specific Media Independent Interfaces (MIIs) allow use of selected Physical Layer devices (PHY) for operation over coaxial, twisted-pair or fiber optic cables. System considerations for multisegment shared access networks describe the use of Repeaters that are defined for operational speeds up to 1000 Mb/s. Local Area Network (LAN) operation is supported at all speeds. Other specified capabilities include various PHY types for access networks, PHYs suitable for metropolitan area network applications, and the provision of power over selected twisted-pair PHY types.*

Ethernet Media Standards

After the publication of the original IEEE 802.3 standard for thick coaxial cable Ethernet, the next development in Ethernet media was the thin coaxial cable variety, inspired by technology first marketed by 3Com Corporation. When the IEEE 802.3 committee standardized the "thin Ethernet" technology (also known as "Cheapernet"), they gave it the shorthand identifier of 10BASE2, as explained later in this chapter.

Following the development of the thin coaxial variety of Ethernet came a steady stream of new media varieties over the years, including the unshielded twisted-pair and fiber optic varieties for the 10 Mb/s system. Next, the 100 Mb/s Fast Ethernet system was created, which also included several varieties of twisted-pair and fiber optic media systems. Following the 100 Mb/s system came the 1 Gigabit, 10 Gigabit, and most recently 40 and 100 Gigabit Ethernet media systems. The media systems were all initially specified as supplements to the main IEEE Ethernet standard.

IEEE Supplements

When the Ethernet standard needs to be changed to add a new media system or other capability, the IEEE develops the new standard as a supplement. The supplement may consist of one or more entirely new sections or "clauses" in IEEE-speak, and may also contain changes to existing clauses in the standard. New supplements to the standard are first evaluated by engineering experts at various IEEE meetings; the supplements must then pass a balloting procedure before being voted into the full standard.

New supplements are given a letter designation when they are created. Once the supplement has completed the standardization process, it becomes part of the base standard and is no longer published as a separate supplementary document. On the other hand, you will sometimes see Ethernet equipment described with the letters of the supplement in which it was first standardized (e.g., IEEE 802.3u may be used as a reference to Fast Ethernet). Table 2-1 lists some of the supplements.

Table 2-1. IEEE 802.3 supplements

Supplement	Description
802.3a-1988	10BASE2 thin Ethernet
802.3c-1985	10 Mb/s repeater specifications
802.3d-1987	FOIRL 10 Mb/s fiber link
802.3i-1990	10BASE-T twisted-pair
802.3j-1993	10BASE-F fiber optic
802.3u-1995	100BASE-T Fast Ethernet and Auto-Negotiation
802.3x-1997	Full-duplex standard
802.3z-1998	1000BASE-X Gigabit Ethernet
802.3ab-1999	1000BASE-T Gigabit Ethernet over twisted-pair
802.3ac-1998	Frame size extension to 1,522 bytes for VLAN tag
802.3ad-2000	Link aggregation for parallel links
802.3ae-2002	10 Gb/s Ethernet
802.3af-2003	Power over Ethernet ("DTE Power via MDI")
802.3ak-2004	10GBASE-CX4 10 Gigabit Ethernet over short-range coaxial cable
802.3an-2006	10GBASE-T 10 Gigabit Ethernet over twisted-pair
802.3as-2006	Frame expansion to 2,000 bytes for all tagging
802.3aq-2007	10GBASE-LRM 10 Gigabit over long-range fiber optic
802.3az-2010	Energy-efficient Ethernet
802.3ba-2010	40 Gb/s and 100 Gb/s Ethernet

The years of formal acceptance of each supplement into the standard are shown. The list is sorted alphabetically, but the years are not all in numeric order. Because of the different rates at which standardization progress was made, the 802.3ac supplement, for example, was adopted into the standard before 802.3ab. Information on the 802.3 supplements and working groups can be found on the Ethernet Working Group's website.

Draft Standards

If you've been using Ethernet for a while, you may recall times when a new variety of Ethernet equipment was being sold while the standard was still in draft form, and before the supplement that described the new variety had been entirely completed or voted on.

This illustrates a common problem: innovation in the computer field, and especially in computer networking, frequently outpaces the more deliberate and slow-paced process of developing and publishing standards.

Vendors are eager to create and market new products, and it's up to you, the customer, to make sure that a product you're considering will work properly in your network system. One way you can do that is to insist on complete information from the vendor as to what version of the standard the product complies with.

It may not be a bad thing if the product is built to a draft version of a new supplement. Draft versions of the supplements can be substantially complete, yet still take months to be voted on by the various standards committees. When buying prestandard equipment built to a draft of the specification, you need to ensure that the draft in question is sufficiently well along in the standards process that no major changes will be made. Otherwise, you could be left out in the cold with network equipment that won't interoperate with newer devices built according to the final published standard.

One solution to this problem is to get a written guarantee from the vendor that the equipment you purchase will be upgraded to meet the final published form of the standard. Note that the IEEE forbids vendors to claim or advertise that a product is compliant with an unapproved draft.

Differences Between DIX and IEEE Standards

When the IEEE developed 802.3 from the original DIX standard, it made some changes in the specifications. One reason for this was that the two groups had different goals. The specifications for the DIX Ethernet standard were developed by the three companies involved, and were intended to describe the Ethernet system—and only the Ethernet system. At the time that the multivendor DIX consortium was developing the first Ethernet standard, there was no open LAN market, nor was there any other multivendor LAN standard in existence. The efforts aimed at creating a worldwide system of open standards had only just begun.

The IEEE, on the other hand, was developing a set of standards intended to integrate into the world of international LAN standards. Consequently, the IEEE made several technical changes required for inclusion in the worldwide standardization effort. The goal was to standardize network technologies under one umbrella, coordinated with the International Organization for Standardization (ISO).[2] The IEEE specifications did permit backward compatibility with early Ethernet systems built according to the orig-

2. According to the ISO website, "Because *International Organization for Standardization* would have different acronyms in different languages (IOS in English, OIN in French for Organisation internationale de normalisation), our founders decided to give it the short form ISO. ISO is derived from the Greek isos, meaning equal. Whatever the country, whatever the language, the short form of our name is always ISO."

inal DIX specifications. Note that this is of historical interest only, though; all Ethernet equipment built since 1985 is based on the IEEE 802.3 standard.

Organization of IEEE Standards

The IEEE standards are organized according to the Open Systems Interconnection (OSI) reference model. This model was developed in 1978 by the International Organization for Standardization. Headquartered in Geneva, Switzerland, the ISO is responsible for setting open, vendor-neutral standards and specifications for items of technical importance.

The ISO developed the OSI reference model to provide a common organizational scheme for network standardization efforts (with perhaps an additional goal of keeping us all confused with reversed acronyms). What follows is a quick, and necessarily incomplete, introduction to the subject of network models and international standardization efforts.

The Seven Layers of OSI

The OSI reference model is a method of describing how the interlocking sets of networking hardware and software can be organized to work together in the networking world. In effect, the OSI model provides a way to arbitrarily divide the task of specifying network behavior into separate chunks, which are then subjected to the formal process of standardization. It's important to remember that OSI is a model for describing network functions, and not an architecture or blueprint for network design.

The OSI reference model describes seven layers of networking functions, as illustrated in Figure 2-1. The lower layers cover the standards that describe how a LAN system moves bits around. The higher layers deal with more abstract notions, such as the reliability of data transmission and how data is represented to the user. The layers of interest for Ethernet are the lowest two layers, Layer 1 and Layer 2, of the OSI model.

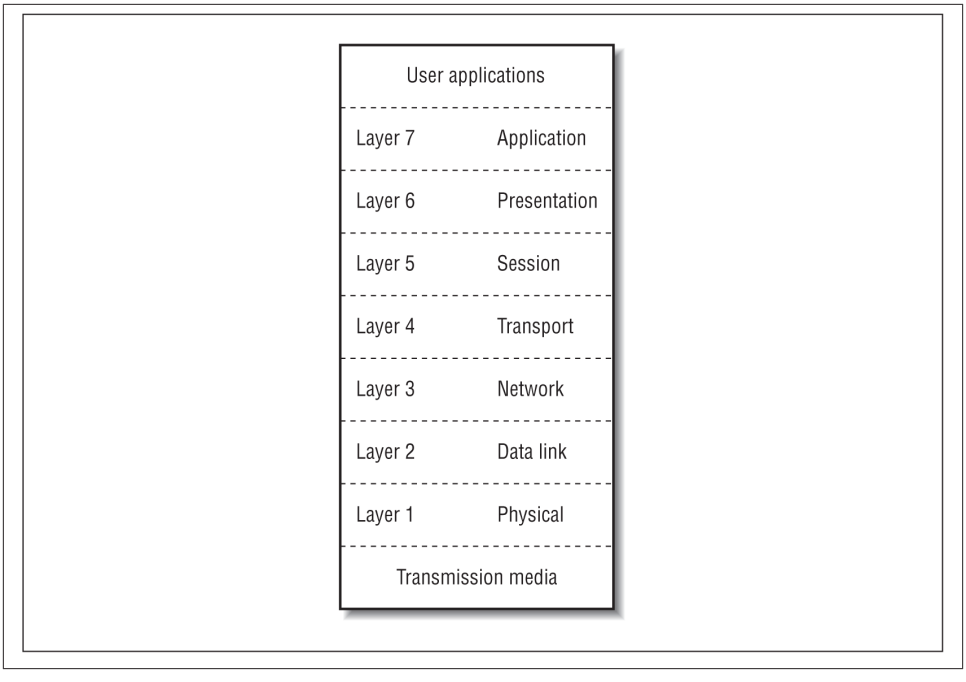


Figure 2-1. The OSI seven-layer model

In brief, the OSI reference model includes the following seven layers, starting at the bottom and working our way to the topmost layer:

Physical layer (Layer 1)
: Standardizes the electrical, mechanical, and functional control of data circuits that connect to physical media.

Data link layer (Layer 2)
: Establishes communication from station to station connected to the same network system. This is the layer that transmits and receives frames and recognizes link addresses. The parts of the Ethernet standard that describe the frame format and Media Access Control protocol belong to this layer.

Network layer (Layer 3)
: Establishes communication from station to station across an internetwork, which is composed of a number of interconnected network systems. This layer provides a level of independence from the lower two layers by establishing higher-level functions and procedures for exchanging data between computers across multiple networks. Standards at this layer of the model describe portions of the high-level network protocols that are carried in the data field of the Ethernet frame. Protocols at and above this layer of the OSI model are independent of the Ethernet standard.

Transport layer (Layer 4)
: Provides reliable end-to-end error recovery mechanisms and flow control, located in the higher-level networking software.

Session layer (Layer 5)
: Provides mechanisms for establishing reliable communications between cooperating applications running on separate computers.

Presentation layer (Layer 6)
: Provides mechanisms for dealing with data representation in applications.

Application layer (Layer 7)
: Provides mechanisms to support end-user applications (e.g., email or web browsers).

IEEE Sublayers Within the OSI Model

The Ethernet standard concerns itself with elements described in Layer 2 (the data link layer) and Layer 1 (the physical layer) of the OSI model. For that reason, you'll sometimes hear Ethernet referred to as a *link layer standard*. To help organize the details of developing specifications for Ethernet, the IEEE defines extra sublayers that fit into the lower two layers of the OSI model, which simply means that the IEEE standard includes some more finely grained layering than the OSI model.

While at first glance these extra layers might seem to be outside the OSI reference model, the OSI model is not meant to dictate the structure of network standards or the design of network products. Instead, the OSI model is an organizational and explanatory tool; sublayers can be added to help deal with the complexity of a given standard.

Figure 2-2 depicts the lower two layers of the OSI reference model and shows how several of the IEEE-specific sublayers are organized. Within the major sublayers shown, there are further sublayers defined for additional MAC functions, new physical signaling standards, and so on. At the OSI data link level (Layer 2), there are IEEE logical link control (LLC) and media access control (MAC) sublayers, which are the same for all varieties and speeds of Ethernet. The LLC layer is an IEEE-defined mechanism for identifying the data carried in an Ethernet frame. The MAC layer defines the protocols used to arbitrate access to the Ethernet system. Both of these sublayers are described in Chapter 3.

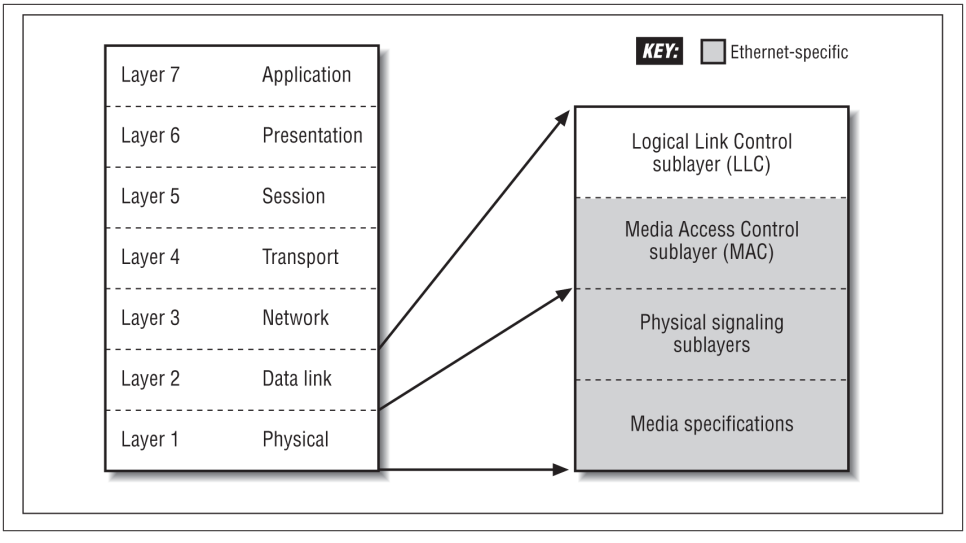


Figure 2-2. The major IEEE sublayers

At the OSI physical layer (Layer 1), you find IEEE sublayers that are specific to the media speed of Ethernet that is being standardized. Each of these sublayers is used to help organize the Ethernet specifications around the functions that must be achieved to make a specific media variety of Ethernet work.

Understanding these sublayers also helps us understand the scope of the standards involved. For example, the MAC portion of the IEEE standard is "above" the lower layer media specifications. As such, the MAC standards are functionally independent of the various physical-layer media specifications, meaning that the MAC sublayer does not change, no matter which physical media variety may be in use.

The IEEE LLC standard is independent of the 802.3 Ethernet LAN standard and does not vary, no matter which LAN system is used. The LLC control fields are intended for use in any LAN system, and not just Ethernet; this is why the LLC sublayer is not formally part of the IEEE 802.3 system specifications.

All of the IEEE sublayers below the LLC sublayer are specific to the individual LAN technology in question, which in this case is Ethernet. To help make this clearer, the Ethernet-specific portions of the standard in Figure 2-2 are shaded, and the LLC sublayer is not shaded.

Below the MAC sublayer, we get into the portions of the standard that are organized in the physical layer of the OSI reference model. The physical layer standards differ depending on the Ethernet media variety in use, and on whether we're describing the original 10 Mb/s Ethernet system, 100 Mb/s Fast Ethernet, 1000 Mb/s Gigabit Ethernet,

10 Gb/s Ethernet, or 40/100 Gb/s Ethernet. These sublayers are described in more detail in Part II.

Levels of Compliance

In developing a technical standard, the IEEE includes only those items whose behavior must be carefully specified to ensure that the system functions correctly. Therefore, all Ethernet interfaces must comply fully with the MAC protocol specifications in the standard to perform their functions identically. Otherwise, the network would not work correctly.

At the same time, the IEEE makes an effort not to constrain the market by standardizing such things as the appearance of an Ethernet interface, or how many connectors it should have on it. The intent is to provide just enough engineering specifications to make the system work reliably and interoperate correctly, without inhibiting competition and the inventiveness of the marketplace. In general, the IEEE has been quite successful in this goal.

Vendor innovation can sometimes lead to the development of devices that are not described in the IEEE standard, and that are not included in the media specifications in the standard. Some of these devices may work well, but they typically will not interoperate with other vendors' equipment because they do not follow the standards.

The Effect of Standards Compliance

How much you should be concerned about all this is largely up to you and your particular circumstances. Another way of saying this is: "Optimality differs according to context."[3] It's up to you to decide how important device and media system interoperability may be, given your particular circumstances (or *context*).

For one thing, not all innovations are bad. After all, the twisted-pair Ethernet media system started life as a vendor innovation that later became a carefully specified media system in the IEEE standard. However, if your goal is maximum predictability and stability for your network given a variety of vendor equipment and traffic loads, then one way to help achieve that goal is by using only equipment that is described in the standard.

One way to decide how important these issues are is to look at the scope and type of network system in question. For an Ethernet that just connects a couple of computers in your house, you may feel that any equipment you can find that helps make this happen at the lowest cost is a good deal. If the equipment isn't described in the official standards,

3. M. A. Padlipsky made this pithy observation about engineering choices in *The Elements of Networking Style* (Englewood Cliffs, NJ: Prentice Hall, 1985), p. 229.

you may not care all that much. In this instance, you are building a small network system, and you probably don't intend for the network to grow very large. The limited scope of your network makes it easier to decide that you are not all that worried about multivendor interoperability.

On the other hand, if you are a network manager of a campus network system, other people using your network will be depending on the network to get their work done. The expanded scope changes your context quite a bit. Campus and enterprise networks always seem to be growing, which makes extending networks to accommodate growth a major priority for you. In addition, network stability under all sorts of traffic loads becomes another important issue. In this very different context, the issues of multivendor interoperability and compliance with the standard become much more important.

IEEE Media System Identifiers

The IEEE has assigned shorthand identifiers to the various Ethernet media systems as they have been developed. The three-part identifiers include the speed, the type of signaling used, and information about the physical medium.

What follows are descriptions of some of the more widely known media systems and their identifiers. These systems include the ones you are most likely to encounter as a network developer or user. There are a number of other systems that have been developed for more specialized environments, such as backplane Ethernet, which are not listed here.

10 Megabit per Second (Mb/s) Media Systems

In the earliest Ethernet media systems, the physical medium part of the identifier was based on the cable distance in meters (m), rounded to the nearest 100 meters. In the more recent media systems, the IEEE engineers dropped the distance convention, and the third part of the identifier, which is indicated by a dash (-), simply identifies the media type used (twisted-pair or fiber optic). In roughly chronological order, the identifiers include the following set:

10BASE5
: This identifies the original Ethernet system, based on thick coaxial cable. The identifier means *10* megabits per second transmission speed, *base*band transmission; the *5* refers to the 500-meter maximum length of a given cable segment. The word *baseband* in this instance means that the transmission medium, thick coaxial cable, is dedicated to carrying one service: Ethernet signals.[4]

4. IEEE 802.3 defines a baseband coaxial system as: "A system whereby information is directly encoded and impressed upon the transmission medium. At any point on the medium only one information signal at a time can be present without disruption." From IEEE Std 802.3-2012, paragraph 1.4.98, p. 22.

10BASE2

Also known as the *thin Ethernet* system, this media variety operates at 10 Mb/s, in baseband mode, with cable segment lengths that can be a maximum of 185 meters in length. If the segments can be at most 185 meters long, why does the identifier say "2," thus implying a maximum of 200 meters? The answer is that the identifier is merely a bit of shorthand and not intended to be an official specification. The IEEE committee found it convenient to round the number up, to keep the identifier short and easier to pronounce. The original version of this lower-cost coaxial Ethernet was nicknamed "Cheapernet."

FOIRL

This stands for *Fiber Optic Inter-Repeater Link*. The DIX Ethernet standard mentioned a point-to-point link segment that could be used between repeaters, but did not provide any media specifications. Later, the IEEE committee developed the FOIRL standard, and published it in 1987. FOIRL segments were originally designed to link remote Ethernet coaxial cable segments together. Fiber optic media's immunity to lightning strikes and electrical interference, as well as its ability to carry signals for long distances, makes it an ideal system for transmitting signals between buildings. The specifications in the original FOIRL segment only provided for linking two Ethernet repeaters together, one at each end of the link. While waiting for a newer set of fiber optic specifications to appear, vendors extended the set of devices that could be connected via fiber, allowing a FOIRL segment to be attached to a station as well. These changes were taken up and added to the newer fiber optic link specifications found in the 10BASE-F standard (described later in this section).

10BROAD36

This system was designed to send 10 Mb/s signals over a broadband cable system. Broadband cable systems support multiple services on a single cable by dividing the bandwidth of the cable into separate frequencies, each assigned to a given service. Cable television is an example of a broadband cable system, designed to deliver multiple television channels over a single cable. 10BROAD36 systems are intended to cover a large area; the 36 refers to the 3,600-meter distance allowed between any two stations on the system.

1BASE5

This standard describes a 1 Mb/s system based on twisted-pair wiring, which did not prove to be a very popular system. 1BASE5 was superseded in the marketplace by 10BASE-T, which provided all the advantages of twisted-pair wiring as well as the higher 10 Mb/s speed.

10BASE-T

The "T" in this identifier stands for "twisted," as in twisted-pair wires. This variety of the Ethernet system operates at 10 Mb/s, in baseband mode, over two pairs of

Category 3 (or better) twisted-pair wires.[5] A hyphen is used in this and all newer media identifiers to distinguish the older "length" designators from the newer "media type" designators.

10BASE-F
: The "F" in this identifier stands for *fiber*, as in *fiber optic media*. This is the 10 Mb/s fiber optic Ethernet standard, adopted as part of the IEEE 802.3 standard in November 1993.

100 Mb/s Media Systems

The identifiers in this category of media systems include the following:

100BASE-T
: This is the IEEE identifier for all 100 Mb/s systems. Because these include both fiber optic and twisted-pair systems, the use of "-T" to describe the whole system is somewhat confusing.

100BASE-X
: This identifier refers to the 100BASE-TX and 100BASE-FX media systems. Both systems are based on the same 4B/5B block signal encoding system, adapted from a 100 Mb/s networking standard called Fiber Distributed Data Interface (FDDI), originally standardized by the American National Standards Institute (ANSI).

100BASE-TX
: This standard describes a Fast Ethernet system that operates at 100 Mb/s, in baseband mode, over two pairs of Category 5 twisted-pair cables. The TX identifier indicates that this is the twisted-pair version of the 100BASE-X media systems. This is the most widely used variety of Fast Ethernet.

100BASE-FX
: This type of Fast Ethernet system operates at 100 Mb/s, in baseband mode, over multimode fiber optic cable.

100BASE-T4
: This variety of Fast Ethernet operates at 100 Mb/s, in baseband mode, over four pairs of Category 3 or better twisted-pair cables. It was not widely deployed and has disappeared from the marketplace.

100BASE-T2
: This standard describes a Fast Ethernet system that operates at 100 Mb/s, in baseband mode, on two pairs of Category 3 or better twisted-pair cables. This variety

5. The category system for classifying cable quality is described in Chapter 15.

was never developed by any vendor, so equipment based on the T2 standard is nonexistent.

1000 Mb/s Media Systems

Common identifiers of 1000 Mb/s media systems include the following:

1000BASE-X
: This is the IEEE identifier for the Gigabit Ethernet media systems based on the 8B/10B block encoding scheme adapted from Fibre Channel, which is a high-speed networking system standardized by ANSI. The 1000BASE-X media systems include 1000BASE-SX, 1000BASE-LX, and 1000BASE-CX. The "X" indicates that they are all based on the same block encoding scheme.

1000BASE-SX
: The "S" in this identifier stands for "short," as in short wavelength. This type of Gigabit Ethernet system uses the short-wavelength fiber optic media segments.

1000BASE-LX
: This variety of Gigabit Ethernet uses the long-wavelength fiber optic media segments.

1000BASE-CX
: This type of Gigabit Ethernet system, based on the original Fibre Channel standard, uses short copper cable media segments.

1000BASE-T
: This is the IEEE shorthand identifier for 1000 Mb/s Gigabit Ethernet over Category 5 or better twisted-pair cable. This system is based on a different signal encoding scheme required to transmit gigabit signals over twisted-pair cabling.

10 Gb/s Media Systems

There are a number of 10 Gb/s media systems, with a few of the most widely used systems specified in the standard described here:

10GBASE-CX4
: 10 Gb/s Ethernet over short-range copper cable assemblies (15 m maximum).

10GBASE-T
: 10 Gb/s Ethernet over unshielded and shielded twisted-pair cables. Category 6A or better twisted-pair cables are required to reach the maximum distance specified.

10GBASE-SR
: 10 Gb/s Ethernet over short-range multimode fiber optic cables.

10GBASE-LR
: 10 Gb/s Ethernet over long-range single-mode fiber optic cables.

40 Gb/s Media Systems

Commonly used 40 Gb/s Ethernet media systems include the following:

40GBASE-CR4
: 40 Gb/s Ethernet over four short-range twinaxial copper cables bundled as a single cable.

40GBASE-SR4
: 40 Gb/s Ethernet over four short-range multimode fiber optic cables.

40GBASE-LR4
: 40 Gb/s Ethernet over four wavelengths carried by a single long-distance single-mode fiber optic cable.

100 Gb/s Media Systems

The most common 100 Gb/s media systems in use today are:

100GBASE-SR10
: 100 Gb/s Ethernet over 10 short-range multimode fiber optic cables.

100GBASE-LR4
: 100 Gb/s Ethernet over four wavelengths carried by a single long-distance single-mode fiber optic cable.

CHAPTER 3

The Ethernet System

An Ethernet network is made up of hardware and software working together to deliver digital data between computers. To accomplish this task, several basic elements combine to make an Ethernet system. This chapter describes these elements, as a familiarity with the basic elements provides a good background for working with Ethernet. We will also see how the Ethernet system is used by high-level network protocols to send data between computers.

This chapter discusses the original half-duplex mode of operation, because that's the system that Ethernet began with. *Half-duplex* simply means that only one computer can send data over the Ethernet channel at any given time. In half-duplex mode, multiple computers share access to a single Ethernet channel by using the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Media Access Control (MAC) protocol. Until the introduction of Ethernet switches, half-duplex was the typical mode of operation for Ethernet devices.

However, these days almost all Ethernet devices are connected directly to a port on an Ethernet switch that is operating in full-duplex mode, and do not share the Ethernet signal channel with other devices. When Ethernet devices are connected to switch ports, the Auto-Negotiation protocol will typically select full-duplex mode, in which the original CSMA/CD protocol is shut off and the two devices on the link can send data whenever they like. Half- and full-duplex mode are both forms of media access control and are described in more detail in Chapter 4.

The Four Basic Elements of Ethernet

The Ethernet system includes four building blocks that, when combined, make a working Ethernet:

- The *frame*, a standardized set of bits used to carry data over the system

- The *Media Access Control protocol*, consisting of a set of rules embedded in each Ethernet interface that allow Ethernet *stations* to access the Ethernet channel, in either half- or full-duplex mode
- The *signaling components*, standardized electronic devices that send and receive signals over an Ethernet channel
- The *physical medium*, the cables and other hardware used to carry the digital Ethernet signals between computers attached to the network

The Ethernet Frame

The heart of the Ethernet system is the frame. The network hardware—which includes the Ethernet interfaces, media cables, and so on—exists simply to move Ethernet frames between computers, or stations. A device connected to the network may be a desktop computer, a printer, or anything else with an Ethernet interface in it. For that reason, the Ethernet standard uses the more general term "station" to describe the networked device, and so will we.

The bits in the Ethernet frame are formed up in specified fields. Figure 3-1 shows the basic frame fields. These fields are described in more detail in Chapter 4.

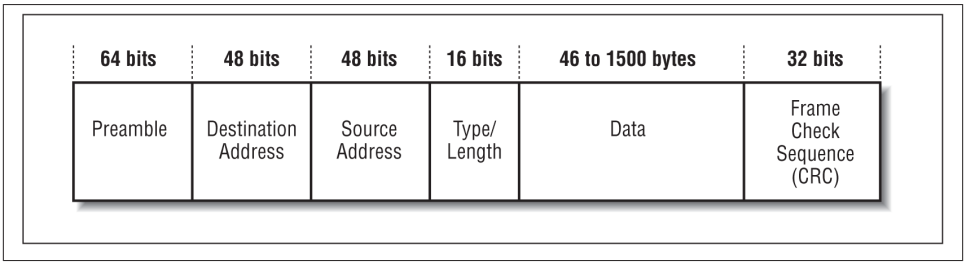


Figure 3-1. An Ethernet frame

Figure 3-1 shows the basic Ethernet frame, which begins with a set of 64 bits called the *preamble*. The preamble was provided to give the hardware and electronics in the original 10 Mb/s Ethernet system some start-up time to recognize that a frame is being transmitted, alerting it to start receiving the data. This is what a 10 Mb/s network uses to clear its throat, so to speak. Newer Ethernet systems running at higher speeds use constant signaling, which avoids the need for a preamble. However, the preamble is still transmitted in these systems to avoid making any changes in the structure of the frame.

Following the preamble are the destination and source addresses. The assignment of one portion of the address is controlled by the IEEE Standards Association (IEEE-SA). When assigning blocks of addresses for use by network vendors, the IEEE-SA

provides a 24-bit *organizationally unique identifier* (OUI).[1] An OUI is assigned to each organization that builds Ethernet interfaces. A manufacturer of Ethernet interfaces creates a unique 48-bit Ethernet address for each interface by using its assigned OUI for the first 24 bits of the address, then assigning the next 24 bits, being careful to ensure that each address is unique.

The resulting 48-bit address is often called the *hardware* (or *physical*) *address*, to make the point that the address has been assigned to the Ethernet interface. It is also called a media access control (MAC) address, because the Ethernet media access control system includes the frame and its addressing.

This allows a vendor of Ethernet equipment to provide a unique address for each interface it builds. Providing unique addresses during manufacturing avoids the problem of two or more Ethernet interfaces on a network having the same address. This also eliminates any need to locally administer and manage Ethernet addresses.

Following the addresses in an Ethernet frame is a 16-bit type or length field. Most often, this field is used to identify what type of high-level network protocol is being carried in the data field (e.g., TCP/IP). This field may also be used to carry length information, as described in Chapter 4.

After the type field can come anywhere from 46 bytes to 1,500 bytes of data. The data field *must* be at least 46 bytes long. This minimum length assured that the frame signals stayed on the network long enough for every Ethernet station in the original 10 Mb/s half-duplex system to hear the frame within the correct time limits. If the high-level protocol data carried in the data field is shorter than 46 bytes, then padding data is used to fill out the data field.

Finally, at the end of the frame there's a 32-bit frame check sequence (FCS) field.The FCS contains a cyclic redundancy check (CRC) that provides a check of the integrity of the data in the entire frame. The CRC is a unique value that is generated by applying a polynomial to the pattern of bits that make up the frame. The same polynomial is used to generate another checksum at the receiving station. The receiving station's checksum is then compared to the checksum generated at the sending station. This allows the receiving Ethernet interface to verify that the bits in the frame survived their trip over the network system intact.

That's basically all there is to an Ethernet frame. Now that you know what an Ethernet frame looks like, you need to know how the frames are transmitted. This is where the set of rules used to govern when a station gets to transmit a frame comes into play. We'll take a look at those rules next.

1. For information on acquiring or looking up an OUI, see Appendix A.

The Media Access Control Protocol

In this section, we will briefly describe the half-duplex mode of operation that is the basis of the original 10 Mb/s Ethernet system. The original system was based on the CSMA/CD protocol, which is a set of rules designed to arbitrate access to a shared channel among a set of stations connected to that channel. Note that today most stations use full-duplex mode, which provides a dedicated channel between the station and a switch port. However, to understand the Ethernet standards and Ethernet operation, it's useful to know how the original half-duplex system functions.

It's still possible for a station to use half-duplex mode when connected to a switch port at 10 and 100 Mb/s over a twisted-pair cable. However, higher-speed media systems support full-duplex mode only.

The way the MAC protocol works is fairly simple. Each Ethernet-equipped device operates independently of all other stations on the network; there is no central controller. Stations attached to an Ethernet cable and operating in half-duplex mode are connected to a signaling channel that (because it is shared), requires the use of the CSMA/CD mechanism to control access.

Full-duplex mode uses a dedicated link between stations that operates in both directions simultaneously, and there is no need to control access to the link.

Ethernet uses a *broadcast delivery* mechanism, in which each frame that is transmitted on the shared channel is heard by every station. While this may seem inefficient, the advantage is that putting the address-matching intelligence in the interface of each station allows the physical medium to be kept as simple as possible. On an Ethernet system, all that the physical signaling and media system has to do is accurately transmit bits to every station; the Ethernet interface in the station does the rest of the work.

Ethernet signals are transmitted from the interface and sent over the shared signal channel to every attached station. To send data in half-duplex mode, a station first listens to the channel, and if the channel is idle, the station transmits its data in the form of an Ethernet frame.

The Ethernet standard uses "frame," although the term "packet" is also used by those who are not trying to be precise in their usage. "Packet" is reserved for describing the data transmitted at Layer 3, the network layer, by those who wish to preserve the distinction between Layer 2 and Layer 3 functions.

As each Ethernet frame is sent over the shared signal channel, or *medium*, all Ethernet interfaces connected to the channel read in the bits of the signal and look at the second field of the frame, which contains the destination address as shown in Figure 3-1. An interface compares the destination address of the frame with its own 48-bit unicast address and any multicast address(es) it has been enabled to recognize. An Ethernet interface whose address matches the destination address in the frame will continue to read the entire frame and deliver it to the networking software running on that computer. All other network interfaces connected to the network will stop reading the frame when they discover that the destination address does not match their own unicast address or an enabled multicast address.

Multicast and broadcast addresses

The Ethernet delivery mechanism also supports multicasting, which is more efficient than sending the same frames to multiple recipients. A *multicast* address allows a single Ethernet frame to be received by a group of stations. An application providing streaming audio and video services, for example, can set a station's Ethernet interface to listen for specific multicast addresses in addition to the built-in unicast (physical) address. This allows for a set of stations to be configured as a multicast group, which is provided with a specific multicast address. A single stream of audio packets sent to the multicast address assigned to that group will be received by all stations in that group.

The *broadcast* address, the 48-bit address of all ones, is a special case multicast address. All Ethernet interfaces that see a frame with this destination address will read the frame in and deliver it to the networking software on the computer.

After each frame transmission, all stations on the shared half-duplex network channel with traffic to send must contend equally for the next frame transmission opportunity. This ensures that access to the shared channel is fair, and that no single station can lock out the others. Fair access to the shared channel is made possible through the use of the MAC algorithm embedded in the Ethernet interface located in each station. The media access control mechanism for shared-channel half-duplex Ethernet uses the CSMA/CD protocol.

The CSMA/CD protocol

The CSMA/CD protocol functions somewhat like a dinner party in a dark room, where the participants hear, but do not see, one another. Everyone around the table must listen

for a period of quiet before speaking (*Carrier Sense*). Once a space occurs, everyone has an equal chance to say something (*Multiple Access*). If two people start talking at the same instant, they detect that fact and quit speaking (*Collision Detection*).

To translate this into Ethernet terms, the Carrier Sense portion of the protocol means that each interface must wait until there is no signal on the shared channel before it can begin transmitting. If another interface is transmitting, there will be a signal on the channel; this condition is called *carrier*.

Historically, a carrier signal is defined as a continuous constant-frequency signal, such as the one used to carry the modulated signal in an AM or FM radio system. There is no such continuous carrier signal in Ethernet; instead, "carrier" in Ethernet simply means the presence of traffic on the network. The term originates from the Aloha radio system, discussed in Chapter 1, from which Ethernet was derived.

All other interfaces must wait until carrier ceases and the signal channel is idle before trying to transmit; this process is called *deferral*. With Multiple Access, all Ethernet interfaces have the same priority when it comes to sending frames on the network, and all interfaces can attempt to access the channel when it is idle.

The next portion of the access protocol is called Collision Detection. Given that every Ethernet interface has equal opportunity to access the Ethernet, it's possible for multiple interfaces to sense that the network is idle and start transmitting their frames at the same time. When this happens, the Ethernet signaling devices connected to the shared channel sense the *collision* of signals, which tells these Ethernet interfaces to stop transmitting. Each of the interfaces will then choose a random retransmission time after which to resend its frames, in a process called *backoff*.

The CSMA/CD protocol is designed to provide fair and efficient access to the shared channel so that all stations get a chance to use the network, and no station gets locked out due to some other station hogging the channel. After every packet transmission, all stations use the CSMA/CD protocol to determine which station gets to use the Ethernet channel next.

Collisions

If more than one station happens to transmit on the Ethernet channel at the same moment, then the signals are said to *collide*. The stations are notified of this event and reschedule their transmissions using a random time interval chosen by a specially designed backoff algorithm. Choosing random times to retransmit helps the stations to avoid colliding again on the next transmission.

It's unfortunate that the original Ethernet design used the word *collision* for this aspect of the Ethernet media access control mechanism. If collisions had been called something else, such as distributed bus arbitration (DBA) events, then no one would worry much about their occurrence. To most ears, the word "collision" sounds like something bad has happened, leading many people to incorrectly conclude that collisions are an indication of network failure and that lots of collisions must mean the network is broken.

The truth is that collisions are absolutely normal events on a half-duplex shared-channel Ethernet, and are simply an indication that the CSMA/CD protocol is functioning as designed. As more computers are added, there will be more traffic, resulting in more collisions as part of the normal operation of a half-duplex Ethernet system. Collisions resolve quickly. For example, the design of the CSMA/CD protocol ensures that the majority of collisions on a 10 Mb/s Ethernet will be resolved in microseconds, or millionths of a second. Nor does a normal collision result in lost data. In the event of a collision, the Ethernet interface backs off (waits) for some number of microseconds, and then automatically retransmits the frame.

Half-duplex mode networks with very heavy traffic loads may experience multiple collisions for each frame transmission attempt. This is also expected behavior. Repeated collisions for a given packet transmission attempt indicate a very busy network. If repeated collisions occur, the stations involved will expand the set of potential backoff times in order to retransmit the data. The expanding backoff process, formally known as *truncated binary exponential backoff* , is a clever feature of the Ethernet MAC protocol that provides an automatic method for stations to adjust to changing traffic conditions on the network. Only after 16 consecutive collisions for a given transmission attempt will the interface finally discard the Ethernet frame. This can happen only if the Ethernet channel is overloaded for a fairly long period of time, or if it is broken.

Hardware

So far, we've seen what an Ethernet frame looks like, and how the CSMA/CD protocol is used to ensure fair access for multiple stations sending their frames over a shared Ethernet channel. The frame and the CSMA/CD protocol are the same for all varieties of Ethernet. Whether the Ethernet signals are carried over coaxial, twisted-pair, or fiber optic cable, the same frame is used to carry the data, and the same CSMA/CD protocol is used to provide the half-duplex shared-channel mode of operation. In full-duplex mode, the same frame format is used, but the CSMA/CD protocol is disabled.

Now that we've seen how the Ethernet frame and MAC protocol work, we next describe Ethernet hardware. There are two basic groups of Ethernet hardware components: the signaling components, used to send and receive signals over the physical medium, and the media components, used to build the physical medium that carries the Ethernet signals. Not surprisingly, these hardware components differ depending on the speed of

the Ethernet system and the type of cabling used. To show the hardware building blocks, we'll look at an example based on the twisted-pair Ethernet media system.

Signaling components

The signaling components for a twisted-pair system include the Ethernet interface located in the computer, a transceiver, and a twisted-pair cable. An Ethernet may consist of a pair of stations linked with a single twisted-pair segment, or multiple stations connected to twisted-pair segments that are linked together with an Ethernet switch. Segments were linked together in the original CSMA/CD half-duplex Ethernet system with signal repeaters, also called repeater *hubs*. Modern Ethernet systems are based on Ethernet switches, which typically function in full-duplex mode.

Figure 3-2 shows two computers (stations) connected to a switch using twisted-pair cables. Each computer contains an Ethernet interface, which provides a connection to the Ethernet system. The interface contains the electronics needed to form up and send Ethernet frames, as well as to receive frames and extract the data from them. The Ethernet interface is typically provided as a set of chips built into the computer's main logic board; in this case, all you will see of the interface is an Ethernet connector on the back of the computer. You can also purchase add-on Ethernet interface cards that plug into a card slot in the computer, or an external interface such as a USB port.

The Ethernet interface connects to the media system using transceiver electronics designed to work with the twisted-pair media. The word "transceiver" is a combination of *trans*-mitter and re-*ceiver*. Most modern desktop and laptop computers include built-in transceivers for twisted-pair connections. A transceiver contains the electronics needed to take signals from the station interface and transmit them to the twisted-pair cable segment, and to receive signals from the cable segment and send them to the interface.

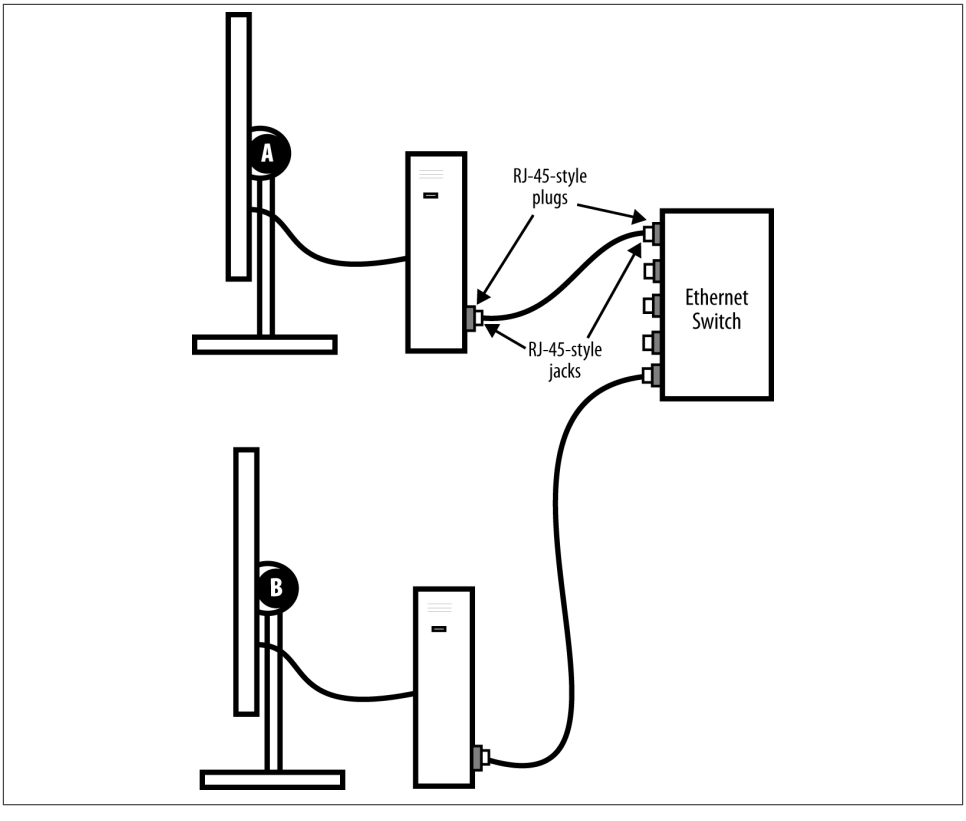


Figure 3-2. A sample twisted-pair Ethernet connection

Prior to the integration of Ethernet electronics in laptops and desktop PCs, it was common to use an external transceiver for both twisted-pair and fiber optic media systems, which connected to a transceiver port on the Ethernet interface. External transceivers are described in Appendix C.

In Figure 3-2, the Ethernet interface in Station A is connected directly to the twisted-pair cable segment, as it is equipped with an internal Ethernet transceiver. The twisted-pair cable uses an eight-pin connector that is also known as an RJ45 plug.

Media components

The cables and signaling components used to create the signal-carrying portion of an Ethernet channel are part of the *physical* medium. The physical cabling components vary depending on which kind of media system is in use. For instance, a twisted-pair cabling system uses different components than a fiber optic cabling system. Just to make

things more interesting, a given Ethernet system may include several different kinds of media systems, all connected together using Ethernet switches.

The design and operation of Ethernet requires that there is only one transmission path between any two stations. An Ethernet grows by extending branches in a network topology called a *tree structure*. A typical network design actually ends up looking less like a tree and more like a complex set of network segments connected to switches throughout wiring closets in a building.

The resulting system of connected cable segments may grow in any direction, and does not have a specified root segment. However, it is essential not to connect Ethernet segments in a loop, as each frame would circulate endlessly until the system was saturated with traffic. On an Ethernet system composed of segments connected with switches, the switches can run a *spanning tree protocol* that automatically detects and shuts off loop paths. The operation of spanning tree is described in Chapter 18.

In the original half-duplex system, an Ethernet LAN consisted of network cable segments linked with one or more signal repeaters. Today, Ethernet systems are built using switches to connect multiple network segments together, with all of the segments usually operating in full-duplex mode. In full-duplex mode, each station has a dedicated connection to the switch port and does not share the Ethernet channel bandwidth on that link with any other computer.

Network Protocols and Ethernet

Now that we've seen how frames are sent over Ethernet systems, let's look at the data being carried by a frame. Data that is being sent between computers is carried in the data field of the Ethernet frame and is structured according to higher-level network protocols. The high-level network protocol information carried inside the data field of each Ethernet frame establishes communications between applications running on computers connected to the network. The most widely used system of high-level network protocols is called the Transmission Control Protocol/Internet Protocol (TCP/IP) suite.

The important thing to understand is that the high-level protocols are *independent* of the Ethernet system. In essence, an Ethernet LAN with its hardware and Ethernet frames is simply a delivery service for data being sent by applications using high-level network protocols. The Ethernet LAN itself doesn't know or care about the high-level network protocol packet being carried in the data field of the Ethernet frame.

Best-Effort Delivery

This is a good place to point out that the Ethernet MAC protocol does not provide a guaranteed data delivery service. Ethernet does not provide strict guarantees for the reception of all data. Instead, the Ethernet MAC protocol makes a "best effort" to deliver

each frame without errors. If bit errors occur during transmission, then a frame may be dropped. Ethernet was designed this way to keep the basic frame transmission system as simple and inexpensive as possible, by avoiding the complexities of establishing guaranteed reception mechanisms at the link layer.

It is assumed that higher layers of network operation, such as TCP/IP, provide the mechanisms needed to establish and maintain reliable data connections when required. Even so, the vast majority of Ethernets operate with very few bit errors or dropped frames. The physical signaling system is designed to provide a very low bit error rate.

The details of how network protocols function are an entirely separate subject from how the Ethernet system works, and are outside the scope of this book. However, the most common use for an Ethernet is to send high-level network protocol packets between computers, so we'll provide a brief example of how high-level network protocols and the Ethernet system work together.

Design of Network Protocols

Network protocols are easy to understand, because we all use some form of protocol in daily life. For example, there's a certain protocol to sending a letter. We can compare the act of sending a letter to what a network protocol does, to see how each works. Sending a letter has a well-known "protocol" that has been standardized through custom. The letter includes a message to the recipient, and the name of the sender. After the letter is composed, it is placed into an envelope, and the name and address of both the recipient and the sender are written on the envelope. The envelope is given to a delivery system, such as the post office, which handles the details of getting the envelope and its contents to the recipient's address. The positions of the addresses and the allowable sizes of the envelopes are also standardized within this "postal protocol," as they are in network protocols.

A network protocol acts much like the letter-sending protocol just described. To carry data between applications, the network software on your computer creates and sends a network protocol packet with its own private data field that corresponds to the message of the letter. The sender's and recipient's names (or protocol addresses) are added to complete the packet. After the high-level network software has created the packet, then the entire network protocol packet is stuffed into the data field of an Ethernet frame. Next, the 48-bit destination and source addresses are added, and the frame is handed to the Ethernet system for delivery to the right computer.

Figure 3-3 shows network protocol data traveling from Station A to Station B. The data is depicted as a letter that is placed in an envelope (i.e., a high-level protocol packet) that has network protocol addresses on it. This letter is stuffed into the data field of an Ethernet frame, shown here as a mailbag. The analogy is not exact, in that each Ethernet frame only carries one high-level protocol "letter" at a time and not a whole bagful, but

you get the idea. The Ethernet frame is then placed on the Ethernet media system for delivery to Station B.

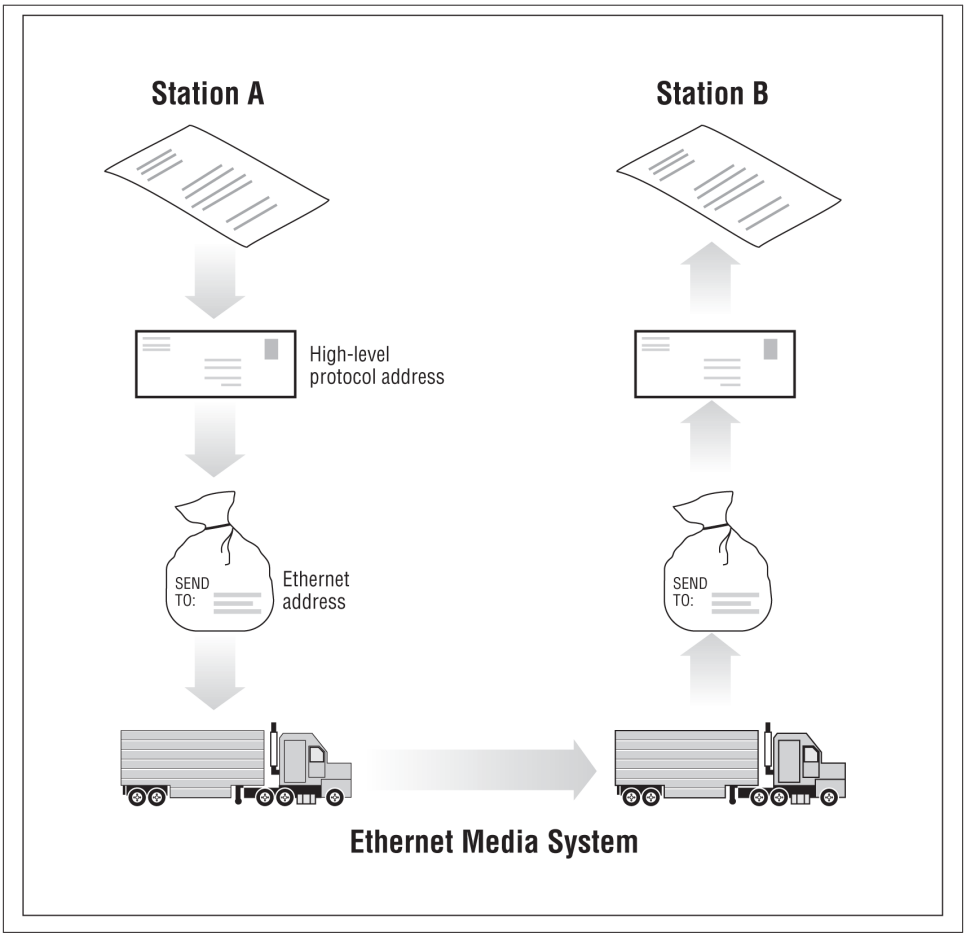


Figure 3-3. Ethernet and network protocols

Protocol Encapsulation

A high-level protocol packet, independent of the Ethernet system, has its own addresses and data. The data field of the Ethernet frame is used to carry the high-level protocol packet. This kind of organization is called *encapsulation*, and it's a common theme in the networking world. Encapsulation is the mechanism that allows independent systems, such as network protocols and Ethernet LANs, to work together.

With encapsulation, the Ethernet frame carries the network protocol packet by treating the entire packet as just so much unknown data, placed into the data field of the Ethernet

frame. Upon delivery of the Ethernet frame to the destination address, it's up to the network software running on that station to deal with the protocol packet extracted from the Ethernet frame's data field.

Just like a trucking system carrying packages, the Ethernet system is fundamentally unaware of what is packed inside the high-level protocol packets that it carries between computers. This allows the Ethernet system to carry all manner of network protocols without worrying about how each high-level protocol works.

In order to get the network protocol packet to its destination, however, the high-level network protocol software and the Ethernet system must interact to provide the correct destination address for the Ethernet frame. When using TCP/IP, the destination address of the IP packet is used to discover the Ethernet destination address of the station for which the packet is intended. Let's look briefly at how this works.

Internet Protocol and Ethernet Addresses

High-level network protocols have their own system of addresses, such as the 32-bit address used by IPv4, which is currently the most widely used version of the Internet Protocol (IP).

The newer IPv6, which provides larger addresses, is being deployed in parallel with the existing IPv4 system.

The Internet Protocol networking software in a given computer is aware of both the 32-bit IP address assigned to that computer, and the 48-bit Ethernet address of its network interface. However, when first trying to send a TCP/IP packet over the Ethernet, it doesn't know what the Ethernet addresses of the other stations on the network are.

To make things work, there needs to be a way to discover the Ethernet addresses of other IP-based computers on the local network. The TCP/IP network protocol system accomplishes this task by using a separate protocol called the Address Resolution Protocol (ARP).

Operation of the ARP protocol

The ARP protocol is fairly straightforward. Figure 3-4 shows two stations, Station A and Station B, sending and receiving an ARP packet over an Ethernet.

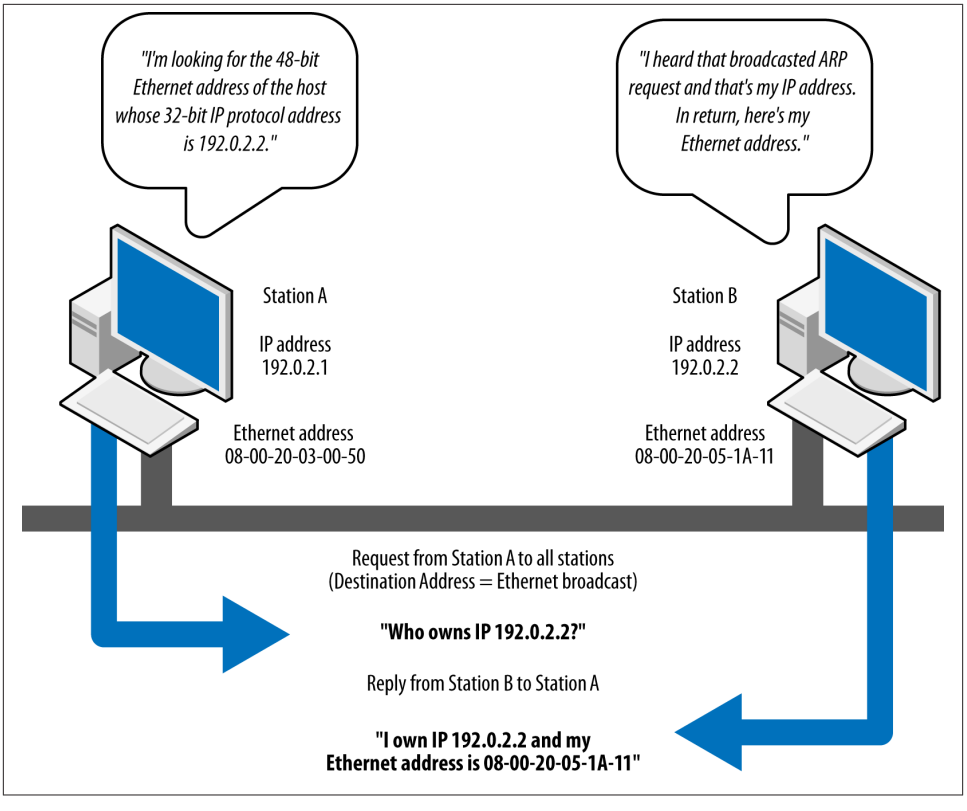


Figure 3-4. Using ARP over an Ethernet

Station A has been assigned the 32-bit IP address of 192.0.2.1, and wishes to send data over the Ethernet system to Station B, which has been assigned IP address 192.0.2.2. Station A sends a packet to the broadcast address, containing an ARP request. The ARP request basically says, "Will the station on this Ethernet that has the IP address of 192.0.2.2 please tell me what the 48-bit hardware address of its Ethernet interface is?"

Because the ARP request is sent in a broadcast frame, every Ethernet station on the same Ethernet LAN reads it and hands the ARP request to the networking software running on that station.

It would have been a better idea to use a specified multicast address for this purpose, so that only IP-speaking computers would receive IP ARPs and computers running other protocols would not be bothered. The address resolution system in the IPv6 version of IP uses multicast addressing for this reason. However, the ARP protocol was one of the first developed for this purpose, and the advantage of using Ethernet multicast addresses was not widely understood by all high-level protocol developers at the time.

Following the broadcast, only Station B with IP address 192.0.2.2 will respond, sending a packet containing the Ethernet address of that station back to the requesting station. Now Station A has an Ethernet address to which it can send frames containing data destined for Station B, and the high-level protocol communication can proceed.

If no station responds, then Station A simply drops the packets destined for 192.0.2.2. Station A will continue to send ARP requests, but in the absence of any response, there is nothing else to do but to discard the packets.

The computers involved build a table in memory called the *ARP cache* to hold the IP addresses and their associated Ethernet addresses. Once this table is created by the ARP protocol, the network software can then look up the IP addresses in the ARP table and find which Ethernet address to use when sending data to a given IP-based machine on the network.

Reaching a station on a separate network

To reach a station that is located on a different IP network, the high-level network software needs to send the packet to a *network router*. Network routers are used to connect network segments together using a high-level network protocol address structure. In the case of TCP/IP, a range of IP addresses are assigned to each separate network. Typically, these IP networks correspond to separate Ethernet LANs connected to the router(s) at a given site.

The high-level network software on a given computer is provided with the local IP address range and the address of at least one router. If the destination address of the packet being sent is not part of the local range, then the software knows that it must send the packet to a router for delivery to a remote network. To do that, an ARP request is made for the router's Ethernet address, in the same way that a station address is requested. Then the Ethernet frame carrying the packet is sent to the router, to be delivered to the remote device.

That's all there is to it. As you can see, the ARP protocol provides the "glue" between the 32-bit addresses used by the IP network protocols and the 48-bit addresses used by the Ethernet interfaces. The two systems operate independently, interacting when there is a need for address discovery.

Looking Ahead

This chapter has provided a brief introduction to the basic elements of an Ethernet system, and how they operate. This description was based on the original half-duplex shared-channel system, which was the primary mode of operation for many years.

However, the development of full-duplex Ethernet and Ethernet switches changed all that, and the most common mode of operation today is full duplex over link segments.

In the next chapter, we discuss the operation of full-duplex Ethernet, along with a detailed description of the Ethernet frame. This provides the knowledge you need to build and manage modern Ethernet systems.

CHAPTER 4

The Ethernet Frame and Full-Duplex Mode

The tutorial in Chapter 3 introduced the Ethernet system and provided a brief look at how it works. In this chapter, we take a more detailed look at the Ethernet frame and the full-duplex mode of operation. You don't need to know all the details of the frame and Ethernet system operation in order to build and use Ethernets. However, an understanding of these elements can certainly help when designing networks or troubleshooting problems.

The original half-duplex mode Media Access Control (MAC) protocol was designed to allow a set of stations to compete for access to a shared Ethernet channel, based on coaxial cable segments linked with signal repeaters. The half-duplex media access control protocol is based on carrier sense with multiple access and collision detection, which gives rise to the CSMA/CD acronym.

The development of full-duplex media systems made it possible for Ethernet links to operate in full-duplex mode, providing a higher-performance mode of operation than the one supported over shared channels using CSMA/CD. The Auto-Negotiation protocol described in Chapter 5 automatically selects the highest-performance mode of operation over a link, typically resulting in full-duplex mode for Ethernet connections. Today, the vast majority of Ethernet links operate in full-duplex mode, which we will describe in this chapter.

However, half-duplex mode is still supported for Ethernet interfaces operating at 10 or 100 Mb/s over twisted-pair cables, and you may find a station connected to a switch port over a link that is in half-duplex mode. The operation of the original half-duplex mode is described in detail in Appendix B.

A twisted-pair link segment is capable of supporting full-duplex operation by virtue of having two pairs of wires that support data being sent in both directions. A station operating in half-duplex mode while it is connected to a twisted-pair media system may indicate a misconfigured link, or an issue with the Auto-Negotiation system. See Chapter 5 for details.

To simplify the description of these elements, this chapter is in two parts. The first two sections look at the structure of the frame and the full-duplex media access control system. The following two sections examine flow control and describe how the high-level network software on a computer uses Ethernet frames to send data.

The Ethernet Frame

The organization of the Ethernet frame is central to the operation of the system. The Ethernet standard determines both the structure of a frame and when a station is allowed to send a frame. The frame was first defined in the original Ethernet DEC-Intel-Xerox (DIX) standard, and was later redefined and modified in the IEEE 802.3 standard. The changes between the two standards were mostly cosmetic, except for the type or length field.

The DIX standard defined a type field in the frame. The first 802.3 standard (published in 1985) specified this field as a length field, with a mechanism that allowed both versions of frames to coexist on the same Ethernet system. Most networking software kept using the type field version of the frame. A later version of the IEEE 802.3 standard was changed to define this field of the frame as being either length or type, depending on usage.

Figure 4-1 shows the DIX and IEEE versions of the Ethernet frame. There are three sizes of frame currently defined in the standard, and a given Ethernet interface must support at least one of them. The standard recommends that new implementations support the most recent frame definition, called an *envelope frame*, which has a maximum size of 2,000 bytes. The two other sizes are *basic frames*, with a maximum size of 1,518 bytes, and *Q-tagged frames* with a maximum of 1,522 bytes.

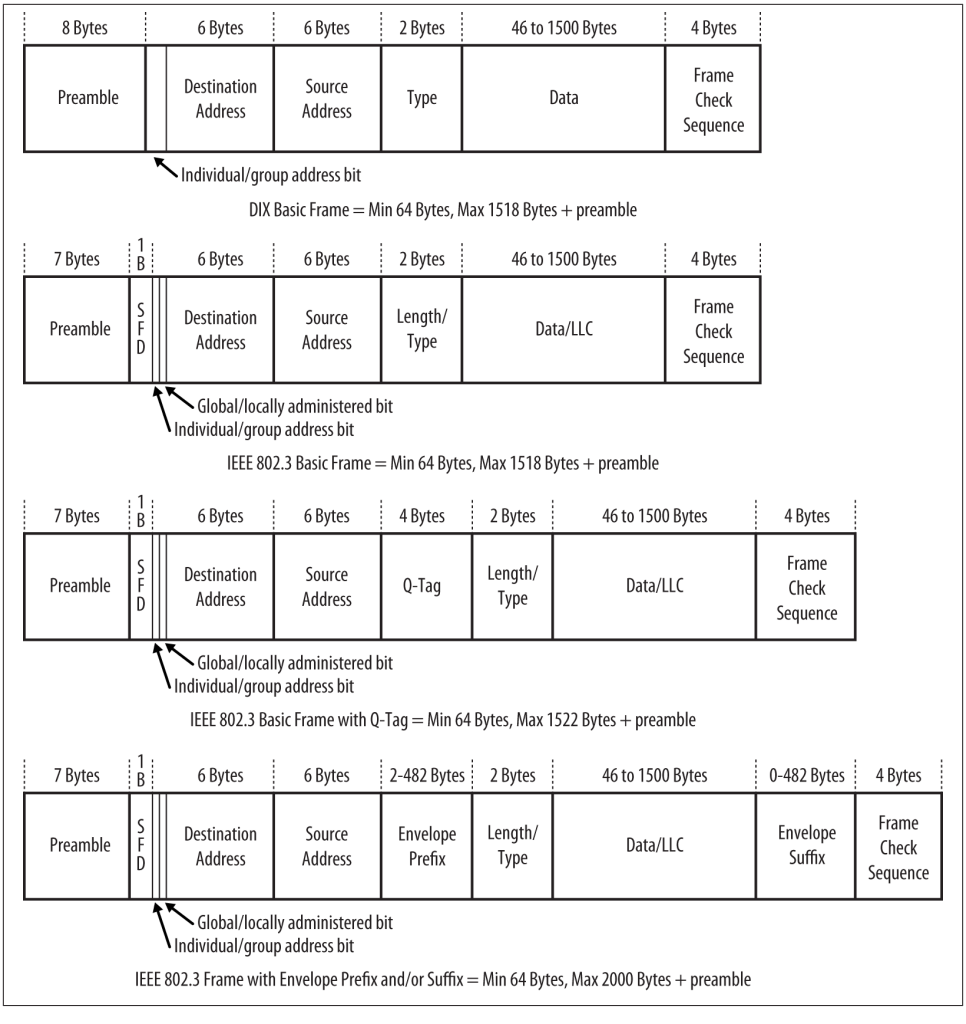


Figure 4-1. DIX Ethernet and IEEE 802.3 frames

Because the DIX and IEEE basic frames both have a maximum size of 1,518 bytes and are identical in terms of the number and length of fields, Ethernet interfaces can send either DIX or IEEE basic frames. The only difference in these frames is in the contents of the fields and the subsequent interpretation of those contents by the network interface software.

Next, we'll take a detailed tour of the frame fields.

Preamble

The frame begins with the 64-bit preamble field, which was originally incorporated to allow 10 Mb/s Ethernet interfaces to synchronize with the incoming data stream before the fields relevant to carrying the content arrived.

The preamble was initially provided to allow for the loss of a few bits due to signal start-up delays as the signal propagates through a cabling system. Like the heat shield of a spacecraft, which protects the spacecraft from burning up during reentry, the preamble was originally developed as a shield to protect the bits in the rest of the frame when operating at 10 Mb/s.

The original 10 Mb/s cabling systems could include long stretches of coaxial cables, joined by signal repeaters. The preamble ensures that the entire path has enough time to start up, so that signals are received reliably for the rest of the frame.

The higher-speed Ethernet systems use more complex mechanisms for encoding the signals that avoid any signal start-up losses, and these systems don't need a preamble to protect the frame signals. However, it is maintained for backward compatibility with the original Ethernet frame and to provide some extra timing for interframe housekeeping, as demonstrated, for example, in the 40 Gb/s system.

While there are differences in how the two standards formally defined the preamble bits, there is no practical difference between the DIX and IEEE preambles. The pattern of bits being sent is identical:

DIX standard
: In the DIX standard, the preamble consists of eight "octets," or 8-bit bytes. The first seven comprise a sequence of alternating ones and zeros. The eighth byte of the preamble contains 6 bits of alternating ones and zeros, but ends with the special pattern of "1, 1." These two bits signal to the receiving interface that the end of the preamble has been reached, and that the bits that follow are the actual fields of the frame.

IEEE standard
: In the 802.3 specification, the preamble field is formally divided into two parts consisting of seven bytes of preamble and one byte called the *start frame delimiter* (SFD). The last two bits of the SFD are 1, 1, as with the DIX standard.

Destination Address

The destination address field follows the preamble. Each Ethernet interface is assigned a unique 48-bit address, called the interface's *physical* or *hardware address*. The desti-

nation address field contains either the 48-bit Ethernet address that corresponds to the address of the interface in the station that is the destination of the frame, a 48-bit multicast address, or the broadcast address.

Ethernet interfaces read in every frame up through at least the destination address field. If the destination address does not match the interface's own Ethernet address, or one of the multicast or broadcast addresses that the interface is programmed to receive, then the interface is free to ignore the rest of the frame. Here is how the two standards implement destination addresses:

DIX standard
: The first bit of the destination address, as sent onto the network medium, is used to distinguish physical addresses from multicast addresses. If the first bit is zero, then the address is the *physical address* of an interface, which is also known as a *unicast address*, because a frame sent to this address only goes to one destination. If the first bit of the address is a one, then the frame is being sent to a *multicast address*. If all 48 bits are ones, this indicates the *broadcast*, or all-stations, address.

IEEE standard
: The IEEE 802.3 version of the frame adds significance to the second bit of the destination address, which is used to distinguish between *locally* and *globally* administered addresses. A globally administered address is a physical address assigned to the interface by the manufacturer, which is indicated by setting the second bit to zero. (DIX Ethernet addresses are always globally administered.) If the address of the Ethernet interface is administered locally for some reason, then the second bit is supposed to be set to a value of one. In the case of a broadcast address, the second bit and all other bits are ones in both the DIX and IEEE standards.

Locally administered addresses are rarely used on Ethernet systems, because each Ethernet interfaces is assigned its own unique 48-bit address at the factory. Locally administered addresses, however, were used on some other local area network systems.

Understanding physical addresses

In Ethernet, the 48-bit physical address is written as 12 hexadecimal digits with the digits paired in groups of two, representing an octet (8 bits) of information. The octet order of transmission on the Ethernet is from the leftmost octet (as written or displayed) to the rightmost octet. The actual transmission order of bits within the octet, however, goes from the least significant bit of the octet through to the most significant bit.

This means that an Ethernet address that is written as the hexadecimal string F0-2E-15-6C-77-9B is equivalent to the following sequence of bits, sent over the Ether-

net channel from left to right: 0000 1111 0111 0100 1010 1000 0011 0110 1110 1110 1101 1001.

Therefore, the 48-bit destination address that begins with the hexadecimal value 0xF0 is a unicast address, because the first bit sent on the channel is a zero.

Source Address

The next field in the frame is the source address. This is the physical address of the device that sent the frame. The source address is not interpreted in any way by the Ethernet MAC protocol, although it must always be the unicast address of the device sending the frame. It is provided for the use of high-level network protocols, and as an aid in troubleshooting. It is also used by switches to build a table associating source addresses with switch ports. An Ethernet station uses its physical address as the source address in any frame it transmits.

The DIX standard notes that a station can change the Ethernet source address, while the IEEE standard does not specifically state that an interface may have the ability to override the 48-bit physical address assigned by the manufacturer. However, all Ethernet interfaces in use these days appear to allow the physical address to be changed, which makes it possible for the network administrator or the high-level network software to modify the Ethernet interface address if necessary.

To provide the physical address used in the source address field, a vendor of Ethernet equipment acquires an organizationally unique identifier (OUI), which is a unique 24-bit identifier assigned by the IEEE. The OUI forms the first half of the physical address of any Ethernet interface that the vendor manufactures. As each interface is manufactured, the vendor also assigns a unique address to the interface using the second 24 bits of the 48-bit address space, and that, combined with the OUI, creates the 48-bit address. The OUI may make it possible to identify the vendor of the interface chip, which can sometimes be helpful when troubleshooting network problems.

Q-Tag

The Q-tag is so called because it carries an 802.1Q tag, also known as a VLAN or priority tag. The 802.1Q standard defines a virtual LAN (VLAN) as one or more switch ports that function as a separate and independent Ethernet system on a switch. Ethernet traffic within a given VLAN (e.g., VLAN 100) will be sent and received only on those ports of the switch that are defined to be members of that particular VLAN (in this case, VLAN 100). A 4-byte-long Q-tag is inserted in an Ethernet frame between the source address and the length/type field to identify the VLAN to which the frame belongs. When a Q-Tag is present, the minimum data field size is reduced to 42 bytes, maintaining a minimum frame size of 64 bytes.

Switches can be connected together with an Ethernet segment that functions as a *trunk* connection that carries Ethernet frames with VLAN tags in them. That, in turn, makes it possible for Ethernet frames belonging to VLAN 100, for example, to be carried between multiple switches and sent or received on switch ports that are assigned to VLAN 100.

VLAN tagging, a vendor innovation, was originally accomplished using a variety of proprietary approaches. Development of the IEEE 802.1Q standard for virtual bridged LANs produced the VLAN tag as a vendor-neutral mechanism for identifying which VLAN a frame belongs to.

The addition of the 4-byte VLAN tag causes the maximum size of an Ethernet frame to be extended from the original maximum of 1,518 bytes (not including the preamble) to a new maximum of 1,522 bytes. Because VLAN tags are only added to Ethernet frames by switches and other devices that have been programmed to send and receive VLAN-tagged frames, this does not affect traditional, or "classic," Ethernet operation.

The first two bytes of the Q-tag contain an Ethernet type identifier of 0x8100. If an Ethernet station that is not programmed to send or receive a VLAN tagged frame happens to receive a tagged frame, it will see what looks like a type identifier for an unknown protocol type and simply discard the frame. VLANs and the contents and organization of VLAN tags are described in Chapter 19.

Envelope Prefix and Suffix

As networks grew in complexity and features, the IEEE received requests for more tags to achieve new goals. The VLAN tag provided space for a VLAN ID and Class of Service (CoS) bits, but vendors and standards groups wanted to add extra tags to support new bridging features and other schemes.

To accommodate these requests, the 802.3 standards engineers defined an "envelope frame," which adds an extra 482 bytes to the maximum frame size. The envelope frame was specified in the 802.3as supplement to the standard, adopted in 2006. In another change, the tag data was added to the data field to produce a *MAC Client Data* field. Because the MAC client data field includes the tagging fields, it may *seem* like the frame size definition has changed, but in fact this is just a way of referring to the combination of tag data and the data field for the purpose of defining the envelope frame.

The 802.3as supplement modified the standard to state that an Ethernet implementation should support at least one of three maximum MAC client data field sizes. The data field size continues to be defined as 46 to 1,500 bytes, but to that is added the tagging information to create the MAC client data field, resulting in the following MAC client data field sizes:

- 1,500-byte "basic frames" (no tagging information)

- 1,504-byte "Q-tagged frames" (1,500-byte data field plus 4-byte tag)
- 1,982-byte "envelope frames" (1,500-byte data field plus 482 bytes for all tags)

The standard notes that:

> *The envelope frame is intended to allow inclusion of additional prefixes and suffixes required by higher layer encapsulation protocols … such as those defined by the IEEE 802.1 working group (such as Provider Bridges and MAC Security), ITU-T or IETF (such as MPLS). The original MAC Client Data field maximum remains 1500 octets while the encapsulation protocols may add up to an additional 482 octets.*[1]

The contents of the tag space are not defined in the Ethernet standard, allowing maximum flexibility for the other standards to provide tags in Ethernet frames. Either or both prefix and suffix tags can be used in a given frame, occupying a maximum tag space of 482 bytes if either or both are present. This can result in a maximum frame size of 2,000 bytes.

The latest standard simply includes the Q-tag as one of the tags that can be carried in an envelope prefix. The standard notes, "All Q-tagged frames are envelope frames, but not all envelope frames are Q-tagged frames." In other words, you can use the envelope space for any kind of tagging, and if you use a Q-tag, then it is carried in the envelope prefix as defined in the latest standard. An envelope frame carrying a Q-tag will have a minimum data size of 42 bytes, preserving the minimum frame size of 64 bytes.

Tagged frames are typically sent between switch ports that have been configured to add and remove tags as necessary to achieve their goals. Those goals can include VLAN operations and tagging a frame as a member of a given VLAN, or more complex tagging schemes to provide information for use by higher-level switching and routing protocols. Normal stations typically send basic Ethernet frames without tags, and will drop tagged frames that they are not configured to accept.

Type or Length Field

The old DIX standard and the IEEE standard implement the type and/or length fields differently:

DIX standard
: In the DIX Ethernet standard, this 16-bit field is called a *type field*, and it always contains an identifier that refers to the *type* of high-level protocol data being carried in the data field of the Ethernet frame. For example, the hexadecimal value 0x0800 has been assigned as the identifier for the Internet Protocol (IP). A DIX frame being used to carry an IP packet is sent with the value of 0x0800 in the type field of the frame. All IP packets are carried in frames with this value in the type field.

1. *IEEE Std 802.3-2012, paragraph 3.2.7, Note 1, p. 56.*

IEEE standard

When the IEEE 802.3 standard was first published in 1985, the type field was not included, and instead the IEEE specifications called this field a *length field*. Type fields were added to the IEEE 802.3 standard in 1997, so the use of a type field in the frame is officially recognized in 802.3. This change simply made the common practice of using the type field an official part of the standard. The identifiers used in the type field were originally assigned and maintained by Xerox, but with the type field now part of the IEEE standard, the responsibility for assigning type numbers was transferred to the IEEE.

In the IEEE 802.3 standard, this field is called a length/type field, and the hexadecimal value in the field indicates the manner in which the field is being used. The first octet of the field is considered the most significant octet in terms of numeric value.

If the value in this field is numerically less than or equal to 1,500 (decimal), then the field is being used as a length field. In that case, the value in the field indicates the number of logical link control (LLC) data octets that follow in the data field of the frame. If the number of LLC octets is less than the minimum required for the data field of the frame, then octets of padding data will automatically be added to make the data field large enough. The content of the padding data is unspecified by the standard. Upon reception of the frame, the length field is used to determine the length of valid data in the data field, and the padding data is discarded.

If the value in this field of the frame is numerically greater than or equal to 1,536 decimal (0x600 hex), then the field is being used as a type field.

The range of 1,501 to 1,535 was intentionally left undefined in the standard.

In that case, the hexadecimal identifier in the field is used to indicate the type of protocol data being carried in the data field of the frame. The network software on the station is responsible for providing any padding data required to ensure that the data field is 46 bytes in length. With this method, there is no conflict or ambiguity about whether the field indicates length or type.

Data Field

Next comes the data field of the frame, which is also treated differently in the two standards:

DIX standard
: In a DIX frame, this field must contain a minimum of 46 bytes of data, and may range up to a maximum of 1,500 bytes of data. The network protocol software is expected to provide at least 46 bytes of data.

IEEE standard
: The total size of the data field in an IEEE 802.3 frame is the same as in a DIX frame: a minimum of 46 bytes and a maximum of 1,500. However, a logical link control protocol defined in the IEEE 802.2 LLC standard may ride in the data field of the 802.3 frame to provide control information. The LLC protocol is also used as a way to identify the type of protocol data being carried by the frame if the type/length field is used for length information. The LLC protocol data unit (PDU) is carried in the first set of bytes in the data field of the IEEE frame. The structure of the LLC PDU is defined in the IEEE 802.2 LLC standard.

The process of figuring out which protocol software stack gets the data in an incoming frame is known as *demultiplexing*. An Ethernet frame may use the type field to identify the high-level protocol data being carried by the frame. In the LLC specification, the receiving station demultiplexes the frame by deciphering the contents of the logical link control protocol data unit. These issues are described in more detail later in this chapter.

FCS Field

The last field in both the DIX and IEEE frames is the frame check sequence (FCS) field, also called the cyclic redundancy check (CRC). This 32-bit field contains a value that is used to check the integrity of the various bits in the frame fields (not including the preamble/SFD). This value is computed using the CRC, a polynomial that is calculated using the contents of the destination, source, type (or length), and data fields. As the frame is generated by the transmitting station, the CRC value is simultaneously being calculated. The 32 bits of the CRC value that are the result of this calculation are placed in the FCS field as the frame is sent. The x^{31} coefficient of the CRC polynomial is sent as the first bit of the field, and the x^0 coefficient as the last.

The CRC is calculated again by the interface in the receiving station as the frame is read in. The result of this second calculation is compared with the value sent in the FCS field by the originating station. If the two values are identical, then the receiving station is provided with a high level of assurance that no errors have occurred during transmission over the Ethernet channel. If the values are not identical, then the interface can discard the frame and increment the frame error counter.

End of Frame Detection

The presence of a signal on the Ethernet channel is known as *carrier*. The transmitting interface stops sending data after the last bit of a frame is transmitted, which causes the

Ethernet channel to become idle. In the original 10 Mb/s system, the loss of carrier when the channel goes idle signals to the receiving interface that the frame has ended. When the interface detects loss of carrier, it knows that the frame transmission has come to an end. The higher-speed Ethernet systems use more complex signal encoding schemes, which have special symbols available for signaling to the interface the start and end of a frame.

A basic frame carrying a maximum data field of 1,500 bytes is actually 1,518 bytes in length (not including the preamble) when the 18 bytes needed for the addresses, length/type field, and the frame check sequence are included. The addition of a further 482 bytes for envelope frames makes the maximum frame size become 2,000 bytes. This was chosen as a useful maximum frame size that could be handled by a typical Ethernet implementation in an interface or switch port, while providing enough room for current and future prefixes and suffixes.

Full-Duplex Media Access Control

The full-duplex mode of operation was added to the standard in 1997, to allow simultaneous communication between a pair of stations over a link. The link between the stations must be composed of a point-to-point media segment, such as twisted-pair or fiber optic media, that provides independent transmit and receive data paths. In full-duplex mode, both stations can simultaneously transmit and receive, which doubles the aggregate capacity of the link. For example, a half-duplex Fast Ethernet twisted-pair segment provides a maximum of 100 Mb/s of bandwidth. When operated in full-duplex mode, the same 100BASE-TX twisted-pair segment can provide a total aggregate bandwidth of 200 Mb/s.

Another major advantage of full-duplex operation is that the maximum segment length is no longer limited by the timing requirements of the original shared-channel half-duplex Ethernet system. In full-duplex mode, the only limits are those set by the signal-carrying capabilities of the media segment. This is especially useful for fiber optic segments, allowing those segments to span long distances.

The full-duplex mode was specified in the 802.3x supplement to the standard. This supplement was approved for adoption into the IEEE 802.3 standard in March 1997. The 802.3x supplement also describes an optional set of mechanisms used for flow control over full-duplex links. The mechanisms used to establish flow control are called *MAC control* and *PAUSE*. First we'll describe how full-duplex mode works, and then we'll show how the MAC control and PAUSE mechanisms can be used to provide flow control over a full-duplex link.

Full-Duplex Operation

The following requirements must be met for full-duplex operation:

- The media system must have independent transmit and receive data paths that can operate simultaneously.
- Exactly two stations can be connected by any full-duplex point-to-point link. There is no contention for use of a shared medium, so the multiple access algorithm (i.e., CSMA/CD) is unnecessary and is not used.
- Both stations on the network link must be capable of, and have been configured to use, the full-duplex mode of operation. This means that both Ethernet interfaces must have the capability to simultaneously transmit and receive frames.

Figure 4-2 shows two stations simultaneously sending and receiving over a full-duplex link segment. The segment provides independent data paths so that both stations can be active without interfering with one another's transmissions.

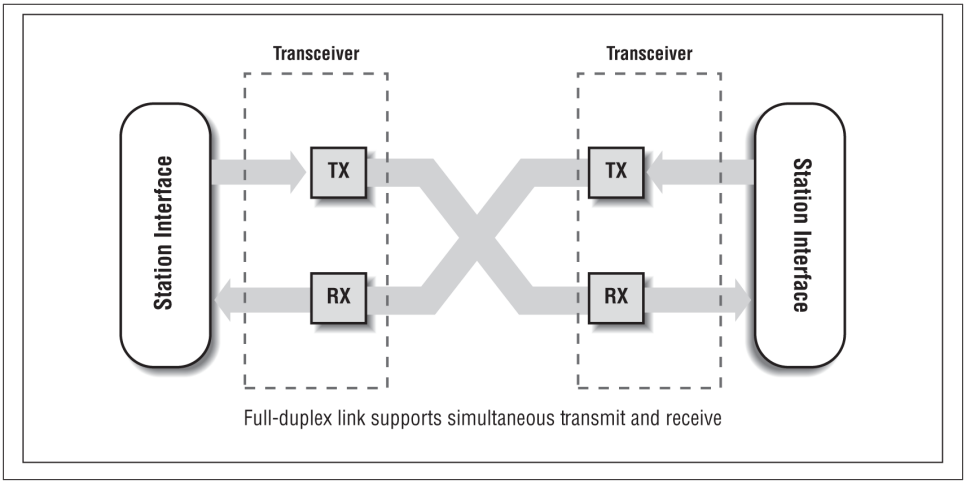


Figure 4-2. Full-duplex operation

When sending a frame in full-duplex mode, the station does not defer to traffic being received on the channel. However, the station still waits for an interframe gap period between frame transmissions, as Ethernet interfaces are designed to expect a gap between successive frames. Providing the interframe gap ensures that the interfaces at each end of the link can keep up with the full frame rate of the link.

A station on a full-duplex link transmits whenever it wishes to, without respect to *carrier sense* (CS), which indicates frames being received from the other station on the receive side of the link segment. There is no *multiple access* (MA), as there is only one station at each end of the link and the Ethernet channel between them is not the subject of access contention by multiple stations. Because there is no access contention, there will

be no collisions either, so the stations at each end of the link also ignore *collision detection* (CD), which indicates frame reception while transmitting.

Effects of Full-Duplex Operation

While full-duplex operation has the potential to double the bandwidth of an Ethernet link segment, it usually won't result in a large increase in performance on a link that connects to a user's computer. That's because few applications send and receive the same amount of data simultaneously. Instead, many applications send some data (e.g., the data resulting from a web click) and then wait for a response. This leads to asymmetric data patterns, in which data that is making requests is sent in one direction, and then larger amounts of data return with the response, often including text, images, or video streams.

On the other hand, full-duplex links between switches in a network backbone system will typically carry multiple conversations between many computers. Therefore, the aggregated traffic on backbone channels will be more symmetric, with both transmit and receive channels seeing roughly the same amount of traffic. For that reason, the largest benefits of a full-duplex bandwidth increase are usually seen in backbone links.

Configuring Full-Duplex Operation

To ensure correct configuration of the Ethernet interfaces at each end of a link, the standard recommends that Ethernet Auto-Negotiation (see Chapter 5) be used whenever possible to automatically configure full-duplex mode. The vast majority of twisted-pair Ethernet interfaces and switch ports support Auto-Negotiation, which will automatically support the highest-performance mode of operation between two stations on a link segment.

It is essential that both ends of a link operating in full-duplex mode are configured correctly, or the link will have data errors. However, using Auto-Negotiation to configure full-duplex operation on a link may not be as simple as it sounds. For one thing, support for Auto-Negotiation is optional for some Ethernet media systems, in which case the vendor is not required to provide Auto-Negotiation capability.

Auto-Negotiation was originally developed for twisted-pair Ethernet devices only, and after the original development of 10BASE-T; thus, it is not supported on all Ethernet media types or older 10BASE-T systems. The older 10 Mb/s and 100 Mb/s fiber optic media systems also do not support the Auto-Negotiation standard, while Gigabit Ethernet fiber optic systems have their own auto-configuration scheme. Therefore, you may find that you have to manually configure full-duplex support on the stations at each end of the link.

On a manually configured link, if only one end of the link is in full-duplex mode and the other is in half-duplex mode, then the half-duplex end of the link will lose frames

due to errors, such as late collisions. Data will still flow across the link, but as the full-duplex end will be sending data whenever it pleases, it will not be obeying the same CSMA/CD rules as the half-duplex end. Because the misconfigured link will still support the flow of data (despite the errors), it is possible that this problem may not be detected right away. Therefore, you need to be aware that this condition can occur, and make sure that both ends of a manually configured link are set for the same mode of operation.

Full-Duplex Media Support

Table 4-1 provides a list of copper Ethernet media systems, and indicates which ones can support the full-duplex mode of operation.

Table 4-1. Full-duplex media support

Media system	Cable type	Full-duplex support?
10BASE5	50 ohm thick coaxial cable	No
10BASE2	50 ohm thin coaxial cable	No
10BASE-T	2-pair twisted-pair	Yes
10BROAD36	75 ohm coaxial cable	No
100BASE-TX	2-pair twisted-pair	Yes
100BASE-T4	4-pair twisted-pair	No
100BASE-T2	2-pair twisted-pair	Yes
1000BASE-SX	2 multimode optical fibers	Yes
1000BASE-LX	2 multimode or single-mode optical fibers	Yes
1000BASE-CX	2-pair shielded twisted-pair	Yes
1000BASE-T	4-pair twisted-pair	Yes
10GBASE-T	4-pair twisted-pair	Yes
10GBASE-CR4	Short-range twinaxial cables	Yes
40GBASE-CR4	Short-range twinaxial cables	Yes

Full-Duplex Media Segment Distances

When a segment is operating in full-duplex mode, CSMA/CD-based MAC operation is disabled. As a result, the cable length limits imposed by the round-trip timing constraints of the CSMA/CD algorithm no longer exist. In the absence of a round-trip timing limit imposed by the CSMA/CD MAC algorithm, the only constraint on cable length is the one imposed by the signal transmission characteristics of the cable. For that reason, some full-duplex segments can be much longer than the same segments operating in half-duplex mode.

For twisted-pair cabling, it is the signal-carrying characteristics of the wires that limit segment length. The 10/100/1000BASE-T and 10GBASE-T media systems have a maximum cabling distance recommendation of 100 meters (328 feet) for twisted-pair cable.

This limit is the same whether the segment is operated in full-duplex or half-duplex mode.

Fiber optic segments, with their excellent signal-carrying characteristics, are mostly limited in length by the timing constraints of half-duplex operation. For that reason, a full-duplex mode fiber optic segment can be considerably longer than the same segment type operating in half-duplex mode. As an example, a 100BASE-FX fiber optic segment using a typical multimode fiber optic cable is limited to segment lengths of 412 meters (1351.6 feet) in half-duplex mode. However, the same media system can reach as far as 2 kilometers (6561.6 feet) when operated in full-duplex mode.

Single-mode fiber optic media can carry signals over longer distances than multimode fiber. Therefore, a full-duplex fiber link can work over considerably longer distances if single-mode fiber is used. In the case of a 100BASE-FX link, single-mode fiber can provide link distances of 20 kilometers (12.42 miles) or more. For full-duplex links, you need to consult the equipment vendor for specifications on the maximum length of the segment.

Ethernet Flow Control

Ethernet flow control is a mechanism that allows an interface or switch port to send a signal requesting a short pause in frame transmission. At the time that this feature was developed, vendors were implementing various approaches to controlling Ethernet frame transmission, in an attempt to manage limited switch and interface resources on busy networks. To provide a vendor-neutral way to signal a request for a brief pause in frame transmission, an explicit flow control message is provided by the optional MAC control and PAUSE specifications in the 802.3x full-duplex supplement.

Today, switch and interface resources are no longer as limited as they once were, and while Ethernet flow control is implemented by vendors, it is not widely used for its original purpose. Instead, you will find PAUSE-based flow control used in data center switch implementations, for example, to provide quality of service for file storage data flows.

The optional MAC control portion of the 802.3x supplement provides a mechanism for real-time control and manipulation of the frame transmission and reception process in an Ethernet station. In normal Ethernet operation, the Media Access Control (MAC) protocol defines how to go about transmitting and receiving frames. In the Ethernet flow control system, the MAC control protocol provides mechanisms to control when Ethernet frames are sent.

The MAC control system provides a way for the station to receive a MAC control frame and act upon it. The operation of the MAC control system is transparent to the normal media access control functions in a station. MAC control is not used for non-real-time functions, such as configuring interfaces, that are handled by network management

mechanisms. Instead, MAC control is designed to allow stations to interact in real time to control the flow of traffic. The specification allows for new functions beyond flow control to be added in the future.

MAC control frames are identified with a type value of 0x8808 (hex). A station equipped with optional MAC control receives all frames using the normal Ethernet MAC functions, and then passes the frames to the MAC control software for interpretation. If the frame contains the hex value 0x8808 in the type field, then the MAC control function reads the frame, looking for MAC control operation codes carried in the data field. If the frame does not contain the 0x8808 value in the type field, then MAC control takes no action, and the frame is passed along to the normal frame reception software on the station.

MAC control frames contain operation codes (*opcodes*) in the data field of the frame. The frame size is fixed at the minimum frame size allowed in the standard, with 46 bytes in the data field. The opcode is contained in the first two bytes of the data field. There is no reliable transport mechanism, so MAC control must be able to deal with the fact that MAC control frames may be lost, discarded, damaged, or delayed.

PAUSE Operation

The PAUSE system of flow control on full-duplex link segments, originally defined in 802.3x, uses MAC control frames to carry the PAUSE commands. The MAC control opcode for a PAUSE command is 0x0001 (hex). A station that receives a MAC control frame with this opcode in the first two bytes of the data field knows that the control frame is being used to implement the PAUSE operation, for the purpose of providing flow control on a full-duplex link segment. Only stations configured for full-duplex operation may send PAUSE frames.

"PAUSE" is not an acronym. Instead, PAUSE is written in uppercase letters to indicate that the word is a formally defined function in the MAC control standard. This is common practice for formally defined words and phrases in the standard.

When a station equipped with MAC control wishes to send a PAUSE command, it sends a PAUSE frame to the 48-bit destination multicast address of 01-80-C2-00-00-01. This particular multicast address has been reserved for use in PAUSE frames. Having a well-known multicast address simplifies the flow control process by making it unnecessary for a station at one end of the link to discover and store the address of the station at the other end of the link.

Another advantage of using this multicast address arises from the use of flow control on full-duplex segments between switches. The particular multicast address used was

selected from a range of addresses reserved by the IEEE 802.1D standard, which specifies basic Ethernet switch (bridge) operation. Normally, a frame with a multicast destination address that is sent to a switch will be forwarded out all other ports of the switch. However, this range of multicast addresses is special—they will not be forwarded by an 802.1D-compliant switch. Instead, frames sent to these addresses are understood by the switch to be frames meant to be acted upon within the switch.

A station sending a PAUSE frame to the special multicast address includes not only the PAUSE opcode, but also the period of pause time being requested, in the form of a two-byte integer. This number contains the length of time for which the receiving station is requested to stop transmitting data. The pause time is measured in units of pause "quanta," where each unit is equal to 512 bit times. The range of possible pause time requests is from 0 through 65,535 units.

Figure 4-3 shows what a PAUSE frame looks like. The PAUSE frame is carried in the data field of the MAC control frame. The MAC control opcode of 0x0001 indicates that this is a PAUSE frame. The PAUSE frame carries a single parameter, defined as the *pause_time* in the standard. In this example, the content of *pause_time* is 2, indicating a request that the device at the other end of the link stop transmitting for a period of two pause quantas (1,024 bit times total).

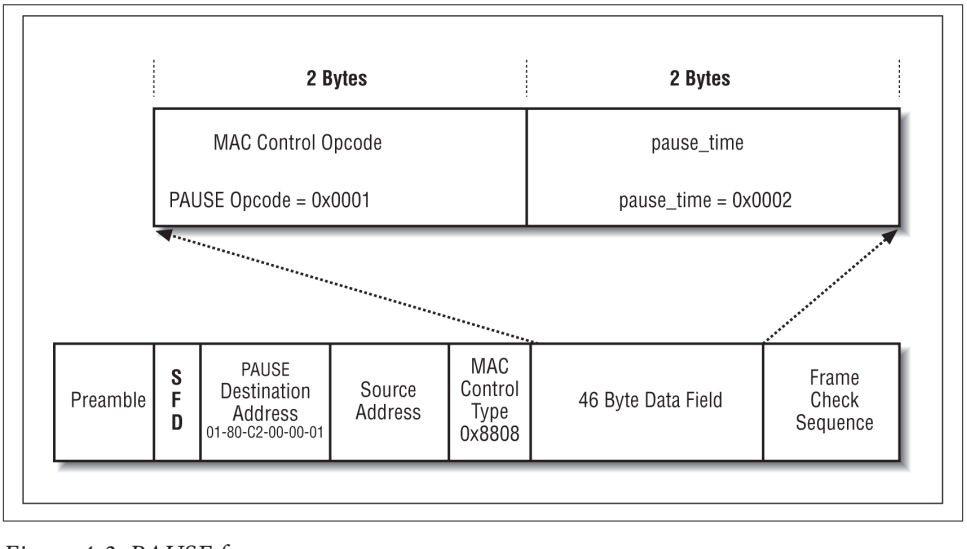


Figure 4-3. PAUSE frame

By using MAC control frames to send PAUSE requests, a station at one end of a full-duplex link can request the station at the other end of the link to stop transmitting frames for a period of time. This provides real-time flow control between switches, or

between a switch and a server that are equipped with the optional MAC control software and connected by a full-duplex link.

High-Level Protocols and the Ethernet Frame

The process of identifying which high-level network protocol data is being carried in the data field of an Ethernet frame is called *multiplexing*. In multiplexing, multiple sources of information can be carried over a single system. In this case, multiple high-level protocols can be sent over the same Ethernet system in separate Ethernet frames.

Multiplexing Data in Frames

The original system of multiplexing for Ethernet is based on using the type field in the Ethernet frame. For example, the high-level protocol software on a computer can create a packet of IP data, and then hand the packet to software that understands how to create Ethernet frames with type fields. The software inserts a hexadecimal value into the type field of the frame; this value corresponds to the type of high-level protocol being carried by the frame. It then hands the data to the interface driver software for transmission over the Ethernet.

The Ethernet interface driver software deals with the details of interacting with the Ethernet interface to send the frame over the Ethernet channel. When carrying IP packets, the type field will be assigned the hexadecimal value 0x0800. The receiving station then uses the value in the type field to identify the protocol data being carried, and thus demultiplex the received frame.

Each layer of the network system is substantially independent from the other layers. Encapsulating the data being passed between layers helps maintain independence between the layers, making it possible for a complex system of network software to be broken down into more manageable chunks. By providing standardized operating system interfaces to the network programmers, the complexity of each network layer is effectively hidden from view.

The programmer is free to write software that hands the completed high-level protocol packet to the appropriate computer system software interface. The details of placing the protocol packet into the data field of an Ethernet frame are automatically dealt with. In this way, an IP-based application, and the IP software itself, can function without major changes regardless of which physical network system the computer happens to be attached to.

Things are made somewhat more complex because of the presence of two methods of identifying data in a frame: one using a type field to identify data, and one using the IEEE 802.2 logical link control (LLC) standard. However, many network drivers are capable of identifying and dealing with multiple frame formats.

IEEE Logical Link Control

As we've seen, the value of the identifier in the length/type field determines which way the field is being used. When used as a length field, the task of identifying the type of high-level protocol being carried in the frame is moved to the 802.2 LLC fields carried in the first few bytes of the data field. Let's look at the LLC fields in a little more detail.

Figure 4-4 shows an IEEE 802.2 LLC protocol data unit, or PDU. The LLC PDU contains a destination service access point (DSAP), which identifies the high-level protocol that the data in the frame is intended for, much like the type field does. After a source service access point (SSAP) and some control data, the actual user data (the data that makes up the high-level protocol packet) follows the LLC fields.

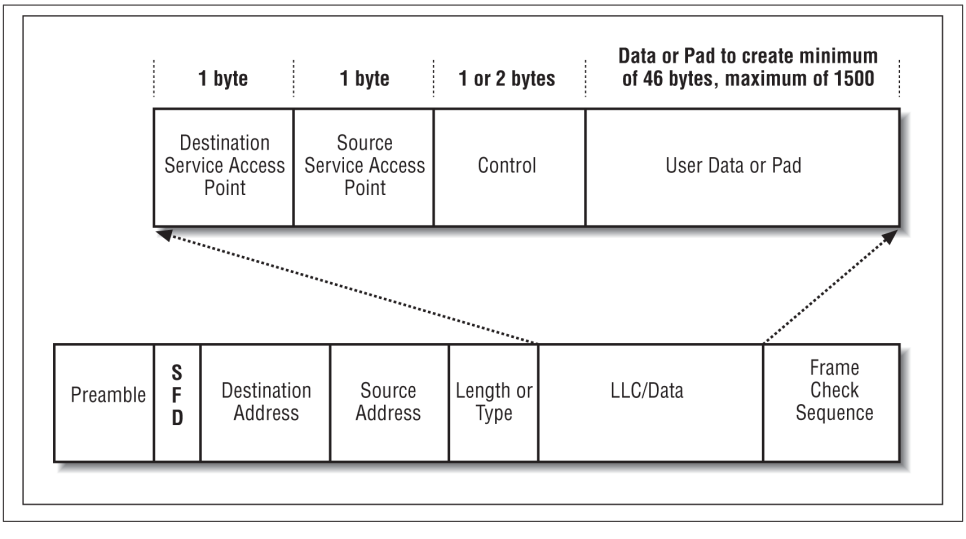


Figure 4-4. LLC PDU carried in an Ethernet frame

Given that TCP/IP uses the type field in Ethernet frames, there is very little use of the IEEE LLC encapsulation on Ethernet systems.

When network protocol software uses the 802.2 LLC fields, multiplexing and demultiplexing work in the same way that they do for a frame with a type field. The difference is that the identification of the type of high-level protocol data is shifted to the DSAP, which is located in the LLC PDU. The whole LLC PDU fits inside the first few bytes of the data field of the Ethernet frame. In frames carrying LLC fields, the actual amount

of high-level protocol data that can be carried is a few bytes less than in frames that use a type field.

You may be wondering why the IEEE went to all the trouble of defining the 802.2 LLC protocol to provide multiplexing when the type field seems to be able to do the job just as well. The reason is that the IEEE 802 committee was created to standardize a set of LAN technologies, and not just the 802.3 Ethernet system. To do that, they needed something that would work no matter which LAN technology was in use.

Because there was no guarantee that all LAN frames would have a type field, the IEEE 802 committee provided the LLC protocol as a method of identifying the type of data being carried by the frame. All LAN systems have a data field, so it is easy enough to write network protocol software that can look at the first few bytes of data in the data field, and then interpret that data in terms of the LLC specifications.

The LLC Sub-Network Access Protocol

Just to make things more interesting, the 802.2 LLC protocol can also be used to carry the original Ethernet type identifiers. In other words, when you send a frame on a non-Ethernet LAN technology that does not provide a type field in its frame, there's a way to use the LLC fields to provide a type identifier. The rationale for this approach comes from the fact that the LLC fields are not large. Given that limitation, the IEEE didn't want to use up the limited number of bits in the LLC fields to provide identifiers for the older high-level protocol types. Instead, a method was created to preserve the existing set of high-level protocol type identifiers, and to reuse them in the IEEE LLC system.

This approach, known as LLC Sub-Network Access Protocol (SNAP) encapsulation, provides yet another set of bytes in the data field of the frame. The contents of the LLC fields of the frame are used to identify another set of bits in the data field, organized according to the SNAP specification, and the SNAP fields are used to carry the older protocol type identifiers. The standard for the use of SNAP encapsulation via IP is documented in RFC 1042. (RFCs can be found at *http://tools.ietf.org*; for more information, see Appendix A.)

If you're writing network protocol software, then SNAP encapsulation is a handy way to continue using the same high-level protocol type identifiers when sending frames over other LAN systems. In the Ethernet system itself, of course, TCP/IP protocol software simply uses the type field, and you don't need to concern yourself with any of this. However, you will probably encounter SNAP encapsulation if you deal with multiple LAN systems at this level of detail.

As a network user, you don't need to lose sleep over which frame format your computers may be using. The choice of frame format is built into your networking software, and there's nothing you need to do about it.

CHAPTER 5
Auto-Negotiation

Automatic configuration of Ethernet interfaces over twisted-pair links and one fiber optic media type is provided by the Auto-Negotiation protocol. Auto-Negotiation is defined in Clause 28 of the Ethernet standard for twisted-pair links, and Clause 37 for the 1000BASE-X fiber optic link. The Auto-Negotiation system ensures that devices at each end of a link can automatically negotiate their configuration to the highest set of common capabilities.

A separate Auto-Negotiation system was developed for use with backplane Ethernet technology, which is defined in Clause 73 of the standard. However, the system defined in Clause 73 is of interest only to developers of Ethernet switches and other devices that use backplane Ethernet technology, and will not be described here.

The need for an automatic configuration system for Ethernet links becomes obvious once you understand that in order to correctly connect a desktop computer to an Ethernet switch port, for example, you must know the speed at which the Ethernet desktop interface and switch port should be set to operate and whether full-duplex mode is supported on the devices at both ends of the link and should therefore be enabled. However, features like the speed and duplex settings are embedded in the network equipment and are invisible to you.

One RJ45 twisted-pair Ethernet port looks a lot like another, and it is not obvious which network options may be supported by the equipment connected to that port. The Auto-Negotiation protocol allows Ethernet equipment to automatically detect and select the correct speed, duplex, and other features, thus relieving you of this configuration task, while ensuring that the Ethernet connection provides the maximum performance supported by the equipment involved in the connection.

The first part of this chapter describes the Auto-Negotiation protocol, and explains how the Auto-Negotiation system works. The rest of the chapter is focused on operational matters, and describes a number of real-world Auto-Negotiation issues that you may encounter. Finally, we show how to develop policies for Auto-Negotiation and link configuration at your site that will result in stable, reliable, and high-performance Ethernet links.

Development of Auto-Negotiation

The specifications for twisted-pair Auto-Negotiation are defined in Clause 28 of the IEEE standard, first published in 1995 as part of the 802.3u Fast Ethernet supplement. Thus, this portion of the standard is sometimes described as "802.3u Auto-Negotiation." The Auto-Negotiation specifications were based on an automatic configuration system called NWay, which was originally developed in the early 1990s by National Semiconductor for use in its isoEthernet system.

The isoEthernet system was designed to provide both a 10 Mb/s Ethernet signal and a separate 6 MHz isochronous channel that could be used to carry voice and video services. Isochronous networking was designed to support applications such as digital voice or video transmission that have strict timing limits that must be met. An isochronous data channel is designed to provide guaranteed bandwidth and signal jitter bounds for a service, permitting data to flow at the rate needed to provide voice or video without audio dropouts or missing portions of video.

The NWay Auto-Negotiation system was developed to discover whether the equipment connected to a given twisted-pair isoEthernet link could support isochronous services, and to enable those services when available. The isoEthernet system and NWay Auto-Negotiation became standardized as the IEEE 802.9 Integrated Services LAN standard. Although the 802.9 standard was not a commercial success, the NWay Auto-Negotiation system was later adopted by IEEE 802.3 to provide automatic negotiation of Ethernet capabilities.

When higher-speed 1000BASE-T and 10GBASE-T Ethernet systems for twisted-pair media were developed, they were designed to also use Clause 28 Auto-Negotiation, so that Clause 28 Auto-Negotiation signals work across all twisted-pair Ethernet media types, including the 10BASE-T, 100BASE-TX, 1000BASE-T, and 10GBASE-T media systems.

Auto-Negotiation for Fiber Optic Media

Auto-Negotiation for fiber optic media segments turned out to be sufficiently difficult to achieve that most Ethernet fiber optic segments do not support Auto-Negotiation. During the development of the Auto-Negotiation standard, attempts were made to develop a system of Auto-Negotiation signaling that would work on the 10BASE-FL and 100BASE-FX fiber optic media systems.

However, these two media systems use different wavelengths of light and different signal timing, and it was not possible to come up with an Auto-Negotiation signaling standard that would work on both. That's why there is no IEEE standard Auto-Negotiation support for these fiber optic link segments. The same issues apply to 10 Gigabit Ethernet segments, so there is no Auto-Negotiation system for fiber optic 10 Gigabit Ethernet media segments either.

The 1000BASE-X Gigabit Ethernet standard, on the other hand, uses identical signal encoding on the three media systems defined in 1000BASE-X. This made it possible to develop an Auto-Negotiation system for the 1000BASE-X media types, as defined in Clause 37 of the IEEE 802.3 standard. The 1000BASE-X Auto-Negotiation system is described later in this chapter.

This lack of Auto-Negotiation on most fiber optic segments is not a major problem, given that Auto-Negotiation is not as useful on fiber optic segments as it is on twisted-pair desktop connections. For one thing, fiber optic segments are most often used as network backbone links, where the longer segment lengths supported by fiber optic media are most effective. Compared to the number of desktop connections, there are far fewer backbone links in most networks. Further, an installer working on the backbone of the network can be expected to know which fiber optic media type is being connected and how it should be configured.

Basic Concepts of Auto-Negotiation

The Auto-Negotiation system makes it possible for Ethernet stations to exchange information about their capabilities over a link segment. This, in turn, allows the stations to perform automatic configuration to achieve the best possible mode of operation over that link. At a minimum, Auto-Negotiation can provide automatic speed matching for Ethernet devices at each end of a twisted-pair link. By using this mechanism, an Ethernet-equipped computer can automatically take advantage of the highest speed offered by a multispeed Ethernet switch port.

The Auto-Negotiation standard includes automatic detection of more than just the speeds of the interfaces at each end of the link. An Ethernet switch capable of supporting full-duplex operation on its ports can advertise that fact using Auto-Negotiation. If a station that also supports full-duplex operation is connected to the switch, then the station and the switch port can automatically configure themselves to use the full-duplex

mode over the link. However, the Auto-Negotiation system does not perform any cable tests, such as detection of the number of wire pairs, or any signal performance tests.

As you'll see later in this chapter, Auto-Negotiation was designed to be extensible and to support the negotiation and configuration of multiple modes of operation and other Ethernet features. Automatic configuration makes it possible for vendors to build twisted-pair Ethernet interfaces that can automatically support several speeds, so that Ethernet switches now support multiple Ethernet speeds on their twisted-pair ports, as do the twisted-pair network interfaces in computers. Depending on the cost of a switch, switch ports may support either 10/100 or 10/100/1000 Mb/s operation. If a switch port supports 10 Gigabit Ethernet, then it typically will support 100/1000/10000 Mb/s operation.

Operation of the Auto-Negotiation system includes the following basic concepts:

Auto-Negotiation operates over link segments
: Auto-Negotiation is designed to work only over a link segment media system. A link segment can have only two devices connected to it—one at each end.

Auto-Negotiation occurs at link initialization
: When a device is turned on, or an Ethernet cable is connected, the link is initialized by the Ethernet devices at each end of the link. Auto-Negotiation and link initialization occur only once, prior to any data being sent over the link. Once the link characteristics are set, they remain the same as long as the link is up.

Twisted-pair Auto-Negotiation uses its own signaling system
: Although each twisted-pair Ethernet media system has a particular method of sending signals over the cable, the twisted-pair Auto-Negotiation system uses its own independent signaling system designed to operate over any twisted-pair cabling that can be used to support Ethernet. If Auto-Negotiation cannot establish a common mode of operation, the link will not come up: if there is no common technology detected at each end of the link, then the Auto-Negotiation protocol will not allow communications. In this case, the switch port or network interface card (NIC) port will not become functional. This is the same situation that occurs when two incompatible Ethernet devices that do not support Auto-Negotiation are connected over a link: the two different systems cannot communicate and the link will not come up.

When describing the operation of Auto-Negotiation, the device at the opposite end of a link from a local device is called the *link partner*. Using the Auto-Negotiation protocol, each device advertises its capabilities to the link partner at the other end of the link. The protocol then selects the highest common denominator between the devices at each end of the link.

Figure 5-1 shows two link segments. Each link segment has two devices: a computer at one end and an Ethernet switch port at the other. Let's assume that computer A only

supports 10/100 operation and computer B only supports 1000BASE-T. Let's further assume that all the ports on the Ethernet switch can support 10/100/1000BASE-T. With Auto-Negotiation, both computers can automatically configure themselves for the highest performance possible over their respective links to the Ethernet switch.

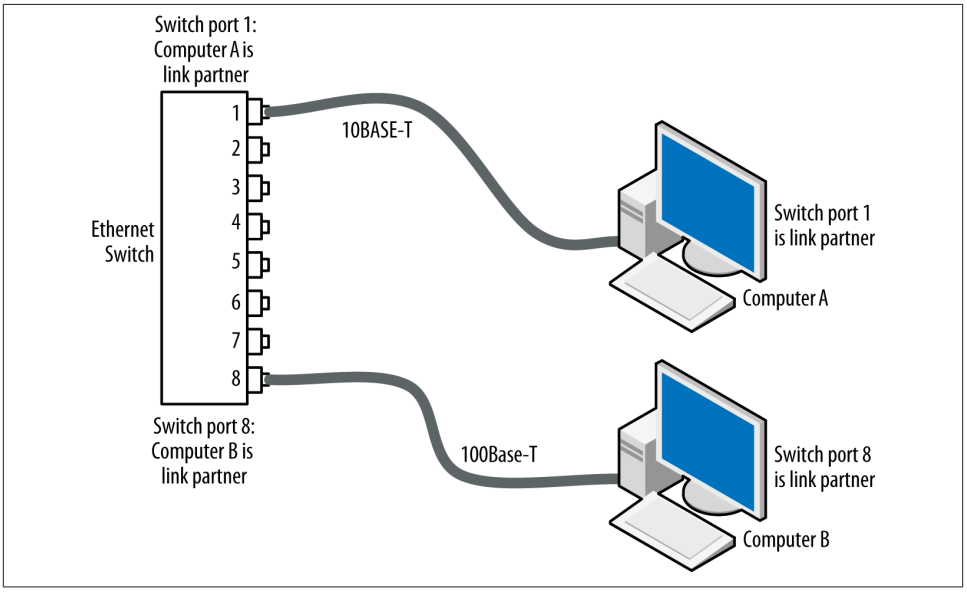


Figure 5-1. Auto-Negotiation link partners

Auto-Negotiation Signaling

The Auto-Negotiation communication method for twisted-pair Ethernet is based on the link integrity test pulse, which was first developed for use in the 10BASE-T media system. The 10BASE-T link pulse was designed to protect a 10 Mb/s twisted-pair Ethernet link segment from an undetected failure in the two pairs of wires used to carry signals over the segment.

To help the Ethernet device at each end of the link avoid the problem of an undetected failure, 10BASE-T transceivers transmit link pulses when there is no data being transmitted. Continual reception of either a data signal or a link pulse assures the interfaces at each end of the link that the link is working correctly.

The signaling methods used in faster Ethernet systems result in a continual stream of signals even when there is no data being sent, which means that the link pulse used in the 10 Mb/s system is not needed to detect a link failure on the faster systems.

The Auto-Negotiation system uses a burst of the *normal link pulse* (NLP) signal, which is called a *fast link pulse* (FLP). The FLP signals are used to carry Auto-Negotiation information between the devices at each end of a twisted-pair link. The Auto-Negotiation FLP signals are specified for transmission over the set of Ethernet twisted-pair media systems, including:

- 10BASE-T
- 100BASE-TX[1]
- 100BASE-T4
- 100BASE-T2
- 1000BASE-T
- 10GBASE-T

Of these media types, the 10BASE-T, 100BASE-TX, and 1000BASE-T systems are very widely used. The 10GBASE-T system is often used in data centers and other environments that require high-speed performance. Equipment based on the 100BASE-T2 standard was never built or sold, and although 100BASE-T4 equipment made it into the marketplace, it too is no longer sold and is rarely found in use.

FLP Burst Operation

When a link with Auto-Negotiation support at both ends is initialized, the devices at each end perform Auto-Negotiation by each sending FLP bursts to their link partner. Link initialization can occur when a link is completed by connecting with a patch cable, when one or both ports of a complete link are manually enabled via their management interface, or when one or both devices at each end of the link are powered on.

Figure 5-2 shows two link partners, an Ethernet station and a switch port, sending FLP bursts to one another over the link. The FLP bursts consist of 33 short pulses, with each pulse 100 nanoseconds (ns) wide. The timing between successive FLP bursts is the same as the timing between NLPs.

1. Shielded twisted-pair cables combined with nine-pin connectors (which are allowed media components for a segment in the 100BASE-TX specifications) are not used in common with any other Ethernet media system, and this segment type does not require or support Auto-Negotiation.

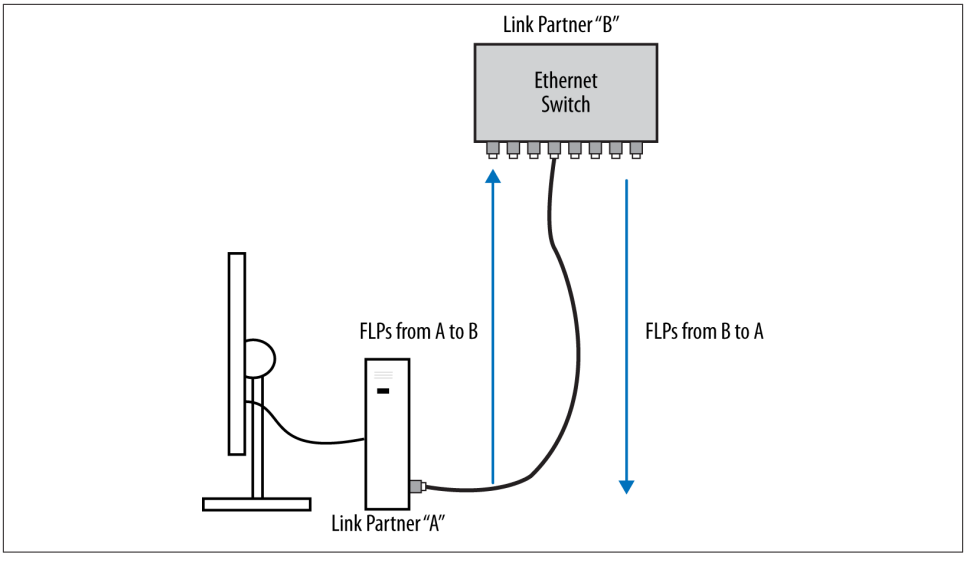


Figure 5-2. The Auto-Negotiation process

The last link pulse in a burst of fast link pulses is timed so as to appear to older 10BASE-T equipment as a normal 10BASE-T link integrity test signal, and the rest of the FLPs are simply ignored by the older equipment. This timing convinces a 10BASE-T device without Auto-Negotiation that it is receiving an NLP, providing backward compatibility with older 10BASE-T equipment that does not support Auto-Negotiation.

Figure 5-3 shows the burst of FLP signals that is used to send information about device capabilities. Of the 33 pulse positions in a fast link pulse burst, the 17 odd-numbered pulse positions each contain a link pulse that represents clock information. The 16 even-numbered pulse positions represent data: the presence of a pulse in an even-numbered pulse position represents a logical 1, and the absence of a pulse represents a logical 0. This encoding scheme is used to transmit the 16-bit *link code words*, or messages, that contain the information used by the Auto-Negotiation protocol.

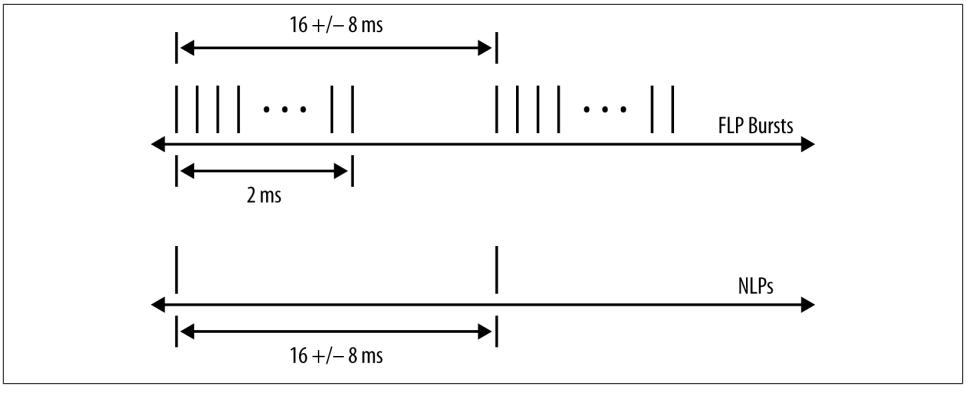


Figure 5-3. Fast link pulses and normal link pulses

When link initialization occurs, the Auto-Negotiation protocol will exchange as many 16-bit messages as are needed. However, many media systems can complete the negotiation in the first message, also called the *base page message*, which is shown in Figure 5-4.

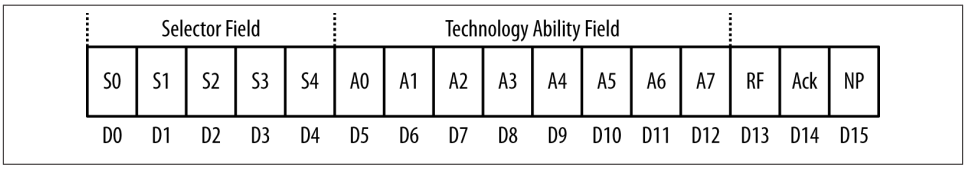


Figure 5-4. Auto-Negotiation base page message

The 16 bits are labeled D0 through D15. Bits D0 through D4 are used as a *selector field* that identifies the type of LAN technology in use, allowing the Auto-Negotiation protocol to be extended to other LAN technologies in the future. For Ethernet, the S0 position of the selector field is set to 1, and all other selector field positions are set to 0.

The 8-bit field from D5 through D12 of the base page message is called the *technology ability field*. The bits in this field are identified as A0 through A7 and are used to indicate support for various technologies, as shown in Table 5-1. If a device supports one or more of these capabilities, it sets the corresponding bit(s) to 1. An auto-negotiating device may use these bits to advertise all of its capabilities in a single base page message.

Table 5-1. Base page technology ability field

Bit	Technology
A0 (D5)	10BASE-T
A1 (D6)	Full-duplex 10BASE-T
A2 (D7)	100BASE-TX

Bit	Technology
A3 (D8)	Full-duplex 100BASE-T
A4 (D9)	100BASE-T4
A5 (D10)	PAUSE operation for full-duplex links
A6 (D11)	Asymmetric PAUSE operation for full-duplex links
A7 (D12)	Reserved for future technology

Bit D13 of the base page message is the *remote fault* bit. This bit may be sent by the remote link partner to indicate that the partner has detected a fault at the remote end. For example, if computer B in Figure 5-1 detected a failure of the incoming signaling, it could set the RF bit to 1 to signal to the switch that the receive side of the link to the computer had failed.

Bit D14 is the *ack* bit, used to acknowledge receipt of the 16-bit message. Negotiation messages are repeatedly sent until the link partner acknowledges them, completing the Auto-Negotiation process for that page or message. After three consecutive messages are received that contain identical information, the link partner sends an acknowledgment. This helps to ensure that the Auto-Negotiation process receives messages correctly even if some bit errors are encountered during the negotiation process.

The final bit in the base page message, D15, is used to signal the next page. Capabilities that are not listed in the base page technology ability field may be advertised in one or more additional next page messages. The next page protocol provides the ability to send vendor-specific commands, or any new configuration commands that may be required as Ethernet evolves. For example, both the 1000BASE-T and 10GBASE-T media systems use the next page protocol to communicate and configure capabilities over the link.

If a device implements this protocol and wishes to send a next page message, it sets the *NP* bit to a value of 1. The next page protocol consists of a two-message sequence. The "message page" indicates the number and type of "unformatted pages" to follow. The unformatted pages contain the data being exchanged between the link partners. Two acknowledgment messages are provided as part of the next page exchange. The first acknowledges receipt of the messages, and the second indicates whether the receiver was able to act upon the information or perform the task defined in the next page message.

Once the stations have completed the Auto-Negotiation process, additional bursts of FLPs will not be sent over the link. The Auto-Negotiation system continuously monitors the link status, and will notice if a link goes down and comes back up (e.g., if a link outage is caused by disconnecting the patch cord to a station). Once a link comes back up, the Auto-Negotiation process will start again from the beginning.

Auto-Negotiation Operation

The Auto-Negotiation protocol contains a set of priorities used by the link partners on a given link to select their highest common set of abilities. When two Auto-Negotiation devices with multiple capabilities are connected, they find their highest common denominator based on a priority table that is specified in the standard. The priority is assigned by technology type, and is not based on the bit ordering in the technology ability field of the base page message.

The priorities defined in the standard for twisted-pair Auto-Negotiation are listed in Table 5-2, ranking from the highest to the lowest priority.

Table 5-2. Auto-Negotiation priority resolution table

Operational mode	Maximum aggregate data transfer rate)
Full-duplex 10GBASE-T	20 Gb/s
Full-duplex1000BASE-T	2 Gb/s
1000BASE-T	1 Gb/s
Full-duplex 100BASE-T2	200 Mb/s
Full-duplex 100BASE-TX	200 Mb/s
100BASE-T2	100 Mb/s
100BASE-T4	100 Mb/s
100BASE-TX	100 Mb/s
Full-duplex 10BASE-T	20 Mb/s
10BASE-T	10 Mb/s

The reasoning behind priority resolution is described in the following excerpt from the standard:

> *The rationale for this hierarchy is straightforward. 10BASE-T is the lowest common denominator and therefore has the lowest priority. Full duplex solutions are always higher in priority than their half-duplex counterparts. 1000BASE-T has a higher priority than 100 Mb/s technologies. 100BASE-T2 is ahead of 100BASE-TX and 100BASE-T4 because 100BASE-T2 runs across a broader spectrum of copper cabling and can support a wider base of configurations. 100BASE-T4 is ahead of 100BASE-TX because 100BASET4 runs across a broader spectrum of copper cabling. The relative order of the technologies specified herein shall not be changed. As each new technology is added, it shall be inserted into its appropriate place in the list, shifting technologies of lesser priority lower in priority. If a vendor-specific technology is implemented, the priority of all IEEE 802.3 standard technologies shall be maintained, with the vendor specific technology inserted at any appropriate priority location.*[2]

2. *IEEE Std 802.3-2012, Annex 28B.3, Section Two, p. 731.*

The full-duplex aggregate data transfer rates in the table reflect the fact that full-duplex operation allows simultaneous two-way transmission, resulting in a maximum aggregate transfer rate for full-duplex operation that is twice the half-duplex transmission rate. All IEEE twisted-pair Ethernet technologies that are listed in the standard are in Table 5-2, whether or not they have been successful in the marketplace.

The 1000BASE-T half-duplex, 1000BASE-T2, and 1000BASE-T4 media systems were not widely deployed or were never sold.

If auto-negotiating devices at both ends of the link can support full-duplex operation, then they will automatically configure themselves to use full-duplex mode. The priority list in Table 5-2 shows that if both devices on the link advertise that they can support 10BASE-T and 100BASE-TX, for example, then the Auto-Negotiation protocol in the link partners will connect using the 100BASE-TX mode instead of 10BASE-T. If both link partners also advertise full-duplex capability, then the 100BASE-TX full-duplex mode will be selected.

If both link partners support the use of PAUSE frames for Ethernet flow control, and if the link is enabled for full-duplex operation, then PAUSE operation will also be enabled at each end of the link. PAUSE operation can be configured via Auto-Negotiation over twisted-pair or Gigabit Ethernet fiber optic segments, but only if the segments are using full-duplex mode. Because the use of PAUSE is independent of data rate or link technology, it is not included in the priority resolution table.[3] The PAUSE flow control system is described in detail in Chapter 4.

If there is no common technology detected at either end of the link, then the Auto-Negotiation protocol will not successfully complete, and the link will not come up. For example, if a device that only supports 10BASE-T is connected to a port on a switch that is configured to only support 100BASE-TX, then no connection will be established on that link. This outcome is unlikely on twisted-pair segments that are configured to allow Auto-Negotiation, as the vast majority of twisted-pair Ethernet switches and NICs support multiple-speed operation on their ports. Switch ports commonly support three twisted-pair media speeds: either 10/100/1000 Mb/s, which is very common, or 100/1000/10,000 MB/s on switches that support the twisted-pair version of 10 Gigabit Ethernet (10GBASE-T).

3. "The use of PAUSE Operation for Full duplex links (as indicated by bits A5 and A6) is orthogonal to the negotiated data rate, medium, or link technology. The setting of these bits indicates the availability of additional DTE capability when full duplex operation is in use…There is no priority resolution associated with the PAUSE operation." From IEEE Std 802.3-2012, Annex 28B.3, Section Two, p. 731.

Parallel Detection

The Auto-Negotiation system includes the ability to work with 10 and 100 Mb/s twisted-pair interfaces that do not support Auto-Negotiation, using a system called Parallel Detection. The Auto-Negotiation feature was designed after the 10 Mb/s twisted-pair media system was developed, and after early 100 Mb/s twisted-pair systems were being sold. Because these varieties of twisted-pair Ethernet predate the Auto-Negotiation standard, support for the Auto-Negotiation protocol is optional in the 10BASE-T and 100BASE-TX media systems.

Given this history, the standards engineers wanted to include the ability for Auto-Negotiation to work with link partners that do not have Auto-Negotiation support. If Auto-Negotiation exists only on one link partner, and if the media speeds involved include 10 and 100 Mb/s Ethernet, then the Auto-Negotiation protocol uses the Parallel Detection system to detect certain characteristics of the non-negotiating partner.

At link initialization time, the absence of fast link pulses from a link partner indicates that the link partner does not support Auto-Negotiation. In that case, if 10BASE-T normal link pulses are present, the NLPs suffice to pass the 10BASE-T link integrity test and to enable 10BASE-T half-duplex operation over the link. Although the 100BASE-TX and 100BASE-T4 Fast Ethernet media systems do not use NLPs, the media signaling characteristics on those systems are sufficiently different to make it possible for Parallel Detection to determine which media system is in use and to bring up the link appropriately.

Auto-Negotiation support is *required* on 1000BASE-TX and 10GBASE-T interfaces to configure essential signal timing characteristics over the link, and it is the only method recognized in the IEEE standard for bringing up those links. Therefore, 1000BASE-T devices must *always* be detectable via Auto-Negotiation, and Parallel Detection is not needed for their operation. Because 100BASE-T4 equipment is not used, in practice this means that Parallel Detection only operates on 10BASE-T or 100BASE-TX links where one link partner is not providing Auto-Negotiation.

Operation of Parallel Detection

Once Parallel Detection determines which 10 or 100 Mb/s media system is being used, then it will set the speed for that particular system. Note that Parallel Detection *always* sets the auto-negotiating device in half-duplex mode. However, if the other end of the link has been manually configured to use full-duplex mode, this can lead to duplex mismatch, which can cause severe performance problems on a link.

To see how Parallel Detection works in practice, let's assume that computer A in Figure 5-1 is a 100BASE-TX device that has been manually configured for speed and duplex. Depending on the vendor and the software used to configure the device, man-

ually configuring the Ethernet interface can result in Auto-Negotiation being shut off entirely, so let's further assume that Auto-Negotiation has been disabled.

As a result, when computer A is powered on, the Ethernet switch in Figure 5-1 will not see any FLPs or NLPs coming from the computer. The Parallel Detection portion of the Auto-Negotiation protocol running on the Ethernet switch port will then detect the type of Fast Ethernet signaling in use, and automatically set the port for 100BASE-TX operation in half-duplex mode. (The standard notes that when Parallel Detection is used, the Auto-Negotiating port must select the half-duplex mode of operation.)[4]

On a number of devices from at least one major switch vendor, manually configuring the speed or duplex on an Ethernet port or NIC will disable Auto-Negotiation. In this case, connecting an auto-negotiating computer to a manually configured switch port results in a link with Auto-Negotiation enabled on just one end.

The Clause 37 specification for 1000BASE-X Auto-Negotiation suggests Auto-Negotiation configuration behavior that will help avoid this issue. The paragraph states:

> *Rather than disabling Auto-Negotiation, the following behavior is suggested in order to improve interoperability with other Auto-Negotiation devices. When a device is configured for one specific mode of operation (e.g. 1000BASE-X Full Duplex), it is recommended to continue using Auto Negotiation but only advertise the specifically selected ability or abilities. This can be done by the Management agent only setting the bits in the advertisement registers that correspond to the selected abilities.*[5]

Parallel Detection and Duplex Mismatch

If a switch port is manually configured for full-duplex mode, and if manual configuration also shuts off Auto-Negotiation on that port, then Parallel Detection will set an auto-negotiating interface connected to that port to half-duplex mode, creating a duplex mismatch on the link that will result in lost frames and low performance. This failure mode is a major reason why you should avoid manual configuration whenever possible, and let Auto-Negotiation handle the task.

The Parallel Detection system is required by the standard to default to half-duplex mode in 10 and 100 Mb/s systems, because an Auto-Negotiating device that is using Parallel Detection has to select some duplex mode, and half-duplex is the safe assumption. Given that 10 and 100 Mb/s equipment must support the original half-duplex mode of Ethernet operation to be compliant with the standard, and given that support for full-duplex mode is optional, selecting half-duplex as the default mode of operation was the only choice left to the developers of the standard.

4. "When selecting the highest common denominator through the Parallel Detection function, only the half-duplex mode corresponding to the selected PMA may automatically be detected." From IEEE Std 802.3-2012, Section Two, Note 2, p. 293.

5. *IEEE Std 802.3-2012, paragraph 37.1.4.4, Section Three, p. 109.*

Table 5-3 shows the results of Auto-Negotiation and Parallel Detection combined with manual configuration for link partners with 10/100 Mb/s capabilities. In Table 5-3, link partner A is not shown but is assumed to have Auto-Negotiation enabled for both speed and duplex. We also assume the worst-case implementation choice on the part of the vendor, in which Auto-Negotiation is disabled for any Ethernet interface that has been manually configured for speed and/or duplex. Link partner B is shown with various configurations, including Auto-Negotiation for both speed and duplex.

Table 5-3. Duplex results of Auto-Negotiation and Parallel Detection

Link partner B Configured speed	Link partner B configured duplex	Result[a]
Auto	Auto	100 FDX[b]
10	HDX	10 HDX
10	FDX	Duplex mismatch
100	HDX	100 HDX
100	FDX	Duplex mismatch

[a] Link partner A is assumed to have Auto-Negotiation enabled for both speed and duplex.

[b] This result assumes that both link partners support 100 Mb/s and full-duplex operation.

On the assumption that manual configuration disables Auto-Negotiation (the worst-case vendor implementation), if one link partner is manually configured for full-duplex operation and the other is Auto-Negotiating, the result will be a duplex mismatch. Even though the Auto-Negotiation protocol on the device at the other end of the link is working correctly, the resulting link will be misconfigured due to the combination of manually configured duplex and lack of Auto-Negotiation on one of the link partners.

Although not shown in the table, if both ends of the link are manually configured (no Auto-Negotiation at either end of the link), and if the manually configured duplex settings at each end do not match, then a duplex mismatch will also result. If manually configured speeds at each end do not match, then there will be no communication possible at all.

Auto-Negotiation Completion Timing

The documentation for one 10/100/1000 Mb/s Ethernet transceiver provided by a major chip manufacturer notes that the processes of Parallel Detection and Auto-Negotiation take approximately 2–3 seconds to complete for 10/100 Mb/s devices, and 5–6 seconds to complete for 1000 Mb/s devices. Auto-Negotiation with next page can take an additional 2–3 seconds to complete, depending on the number of next pages that are sent. These are typical timings, which provide a rough idea of the amount of time that you can expect Auto-Negotiation to require.

By default, most vendors ship switches and Ethernet interfaces with Auto-Negotiation enabled. Leaving Auto-Negotiation enabled on all ports connecting to desktop com-

puters is an effective way to avoid problems with duplex mismatch. However, it is still possible to encounter equipment that does not support Auto-Negotiation or that may have software bugs that may require disabling Auto-Negotiation. Auto-Negotiation debugging is described later in this chapter.

To sum up, modern Ethernet devices support multiple speeds and modes of operation, and are therefore more complex than the original system. The Auto-Negotiation protocol makes it possible for link partners to automatically discover and configure their highest common mode of operation, despite all of the possible speeds and options.

However, Auto-Negotiation may not work successfully if the device at one end of the link is manually configured for full-duplex mode, and if that configuration disables Auto-Negotiation support on the device. In that case, if Auto-Negotiation is enabled on the device at one end of the link and manual full duplex is configured at the other end, the result will be a duplex mismatch that typically causes high frame error rates and lost packets. This can be avoided if the vendor provides an implementation of Auto-Negotiation that continues to operate when speed and duplex settings are manually configured.

Auto-Negotiation and Cabling Issues

The Auto-Negotiation system is designed so that a link will not become operational until matching capabilities exist at each end. However, the Auto-Negotiation protocol is not able to test the quality of the cable used on the link. Therefore, it is up to you to make sure that the correct cable type is in place.

Given a link with an Ethernet switch port at one end, a computer at the other end, and Auto-Negotiation in operation on both devices, and given that both devices support 10, 100, and 1000 Mb/s operation, let's see what effects a difference in cable quality can cause. Assuming that Category 3 cable is used in this link with all four wire pairs connected, then auto-negotiating the 1000BASE-TX mode of operation over this link would be a problem, given that Gigabit Ethernet requires higher-performance Category 5 or 5e cable.[6]

When power is applied or a connection is first made, the switch port and the computer will use Auto-Negotiation to determine the device capabilities at each end of the link connected over Category 3 cable. Auto-Negotiation will choose to operate at the highest-performance mode the devices have in common, which in this example is 1000BASE-TX.

Recall that the Auto-Negotiation link pulses are simply bursts of the same pulses used in 10BASE-T. Those pulses will successfully travel over Category 3 cable, because

6. The Category system for classifying cable quality is described in Chapter 15.

10BASE-T signals were designed to work over Category 3 cable. As a result, the negotiation process will operate correctly, leaving the link configured for 1000BASE-TX operation. However, once the Auto-Negotiation protocol is finished, the signaling switches over to the higher-speed 1000BASE-TX data rate, which requires the use of Category 5 or 5e cable. This means that this link may either operate marginally with a high rate of errors, or not at all.

Structured cabling systems installed in the last several years should be based on Category 5e cabling or better, which avoids this issue. However, Category 5e cabling was not available when the 10BASE-T standard was first developed in the late 1980s, and older cabling plants designed to support 10 Mb/s Ethernet systems may be based on the lower-quality Category 3 cabling.

While Auto-Negotiation is an extremely valuable feature that allows the highest-performance mode to be selected automatically on a given link, it still requires that the correct cable type be in place to support the highest speed mode that may be selected. It's up to you to ensure that the correct cable is in use. If the cabling systems at your site use Category 5, 5e, or better cabling and components, then this is not an issue for the most commonly supported speeds. Assuming that your stations support 10, 100, and 1000 Mb/s, then all three speeds will work over Category 5, 5e, or better cable.

Limiting Ethernet Speed over Category 3 Cable

If your Ethernet links use lower-quality Category 3 cabling, then you may have to manually set the mode of operation. By manually setting the speed, you can make sure that a link does not negotiate a mode of operation that exceeds the capabilities of the cabling for that link. However, if you do so, you will want to ensure that Auto-Negotiation has not been disabled by the manual configuration and continues to function. This can be difficult to determine, as there is no easy way to test for the presence of Auto-Negotiation signals at link initialization. One option is to examine the output of "show" commands on the switch, which may provide information on a port's Auto-Negotiation status.

If the vendor has chosen to disable Auto-Negotiation when speed and/or duplex are manually configured, then you must be very careful to manually set the duplex mode correctly on *both* the switch port and the device connected to the port. As devices are moved and added in a network system, people often forget to maintain the correct manual configuration, which will cause the connection to fail.

A more automatic solution that avoids the need for manual configuration of all devices is for the vendor of the switch to provide a speed-limited Auto-Negotiation setting that does not negotiate above 10BASE-T. One example of this configuration is achieved by setting the ports of a switch to "speed 10baset auto". In this case, the "auto" in the command indicates that the Auto-Negotiation protocol is enabled and is able to negotiate the duplex setting correctly, setting full-duplex operation at both ends when both link partners support it or making sure that both link partners are in half-duplex mode.

According to the switch documentation used for this example, the "speed 10baset auto" setting means that the speed capability that is advertised will be limited to 10 Mb/s, which will ensure that the Auto-Negotiation system does not negotiate a speed that exceeds the capabilities of the cabling. This approach is available in some vendors' equipment. If you have Category 3 cabling that needs to be protected from speeds above 10 Mb/s, you should use this option. If your vendor does not provide such an option, then perhaps it's time to find another vendor.

Cable Issues and Gigabit Ethernet Auto-Negotiation

The 1000BASE-T system of Gigabit Ethernet over twisted-pair requires four twisted pairs of cabling rated at Category 5/5e or better. The Auto-Negotiation system, on the other hand, is designed to work using only two pairs of cabling. As a result, should a pair of auto-negotiating multispeed 10/100/1000 Mb/s interfaces be connected together over a two-pair link, the Auto-Negotiation system would attempt to establish the highest common denominator of performance, which would be 1000BASE-T operation over the link. Given that the 1000BASE-T signaling cannot function over two pairs of cabling, the link would not pass the signal training phase of 1000BASE-T and would not become operational.

Manufacturers of Ethernet chips have developed Ethernet transceivers that will respond to repeated failed attempts to bring up a 1000BASE-T link by automatically downgrading the highest speed capability advertised to the next lower speed. While this feature is not defined in the standard, and is therefore not required to be present, it is a useful capability that can help avoid problems.

In one such transceiver, three failed attempts to establish a 1000BASE-T link will result in an automatic downgrade to 100BASE-T as the highest capability advertised by Auto-Negotiation. If the link is later renegotiated, it will revert to the highest performance advertisements and begin negotiations with 1000BASE-T as its highest advertised capability. This makes it possible for the link to restore operations at 1000 Mb/s, should, for example, an incorrect two-pair patch cable later be replaced with the correct four-pair patch cable required for 1000BASE-T operation.

Crossover Cables and Auto-Negotiation

When connecting a twisted-pair link between two devices, the transmit data from one device must be connected to the receive data on the other device and vice versa. This is called *signal crossover*. The signal crossover can happen inside a switch port, in which case the port may be marked with an "X" to indicate that you can connect a *straight-through* cable between the devices and the signal crossover will be dealt with inside the port.

This approach was used on early switches. However, with the development of auto MDI-X capability (described next), this form of signal path management is no longer used on most switches.

In 1000BASE-T, signals are sent on all four pairs of wires in the cable in both directions simultaneously. To ensure that all of the signals end up on the correct cable pairs in 1000BASE-T, an automatic system to manage the signal locations was developed and standardized in Clause 40 of the IEEE 802.3 standard. This made it possible to use either a straight-through or a crossover cable and let the link partners automatically configure which wire pairs carry the transmit and receive signals to achieve the correct signal paths, in a system called *automatic MDI/MDI-X*.

Unfortunately, when manually configuring port speed, some vendors will disable Auto-Negotiation *and* shut off automatic MDI-X. While the 1000BASE-T mode will continue to work, because MDI-X is part of the 1000BASE-T standard, the 10/100 Mb/s modes of operation can fail due to the lack of MDI-X. As a result, manually configuring the speed of a port could result in the failure of a 10/100 Mb/s link because both Auto-Negotiation and MDI-X are disabled.

For example, a straight-through signal path on all cables that would work fine on a 10/100 Mb/s link with MDI-X enabled will now fail because there is no ability to manage signal paths in the link. This can be very confusing to troubleshoot because the link used to work, and now suddenly it has stopped working, just because the speed configuration was changed. If you encounter this situation, try re-enabling Auto-Negotiation to ensure that MDI-X is also enabled. Note that this is not a failure of Auto-Negotiation or MDI-X. Instead, it is the outcome of a poor implementation decision by the switch vendor.

1000BASE-X Auto-Negotiation

The 1000BASE-T twisted-pair Gigabit Ethernet system operates over copper cabling and uses the same Auto-Negotiation system used by all other twisted-pair Ethernet systems. However, the 1000BASE-X Gigabit Ethernet system has its own system of Auto-Negotiation, defined in Clause 37 of the IEEE 802.3 standard, which operates over 1000BASE-X media segments.

The 10 Mb/s, 100 Mb/s, and 10 Gb/s fiber optic Ethernet media systems use different signaling schemes and operate over different wavelengths of light, so it is not possible to send common Auto-Negotiation signals that all systems could detect. Therefore, Auto-Negotiation is not supported on these media systems.

The designers of 1000BASE-X decided to develop an Auto-Negotiation system that was specific to the three media segment types defined in the 1000BASE-X standard. These systems are 1000BASE-SX and 1000BASE-LX fiber optic segments, and 1000BASE-CX short copper segment. Because all of these systems use the same signal encoding mechanisms, it was possible to use certain signals to send Auto-Negotiation data across all 1000BASE-X systems.

The 1000BASE-X fiber optic media types operate only at 1000 Mb/s so there is no need to automatically negotiate the speed. Further, because no vendors support half-duplex Gigabit Ethernet operation, the full-duplex mode of operation is the only duplex capability advertised by Auto-Negotiation on 1000BASE-X equipment. The only optional capability that is left to be negotiated is support for flow control PAUSE frames.

Note that the 1000BASE-X Auto-Negotiation standard does not include any Parallel Detection capability. In other words, if one link partner is configured to use Auto-Negotiation and the other link partner is not sending Auto-Negotiation signals, then the link will not come up. If the auto-negotiating link partner does not receive Auto-Negotiation signals, there is no fallback to Parallel Detection, and the link will not come up.

Confusingly, the link partner *without* Auto-Negotiation will bring up its end of the link, and the link light will be lit. Because Auto-Negotiation on 1000BASE-X segments uses the same signaling that the 1000BASE-X media system uses, the non-negotiating link partner will see what looks like a normal stream of 1000BASE-X signals coming from the Auto-Negotiating device at the other end of the link, causing it to bring up its end of the link.

However, the auto-negotiating link partner will *not* bring up its end of the link, because it will not see any Auto-Negotiation information from the non-negotiating device. The correct configuration of a 1000BASE-X segment is simple, as long as you know that for the link to come up either both devices on a 1000BASE-X link must be set for Auto-Negotiation, or both must be manually configured with identical settings.

Auto-Negotiation Commands

The Auto-Negotiation protocol is defined in the 802.3 standard, but as we've seen, the Auto-Negotiation implementation and the set of management commands supported on various devices are not standardized. Instead, each vendor is free to implement the management interface and command set for its Ethernet devices as it sees fit. The result is a variety of Auto-Negotiation commands on different vendors' equipment. Just to keep things interesting, different models of switches and other devices from the same vendor may have different management commands, with different command syntax and different outcomes on the operation of Auto-Negotiation.

Auto-Negotiation is performed automatically when the equipment is powered on, or when a link is disconnected and reconnected. Auto-Negotiation can also be triggered manually at any time through the management interface to an Auto-Negotiation device. The most common way to trigger Auto-Negotiation is to use the management interface to manually toggle a switch port to off and back to on, which causes the link to re-initialize.

The command for toggling a port off and then back on varies depending on the vendor, and may require setting the port to "disable" and then "enable," or setting a port interface to "shutdown" and then "no shutdown."

Disabling Auto-Negotiation

It's important to understand that equipment from some vendors will silently disable duplex Auto-Negotiation on an interface or port when the duplex or speed is manually set. Ideally, a warning would be generated to let you know that Auto-Negotiation has been disabled. For a manually configured full-duplex setting, this should include a caution that connecting the manually configured interface to a link partner with Auto-Negotiation enabled can result in a duplex mismatch and very poor performance. Unfortunately, this kind of warning has not been adopted, leading to the potential for confusion as to when Auto-Negotiation is enabled or disabled.

To make things more complex, if you want to manually configure the duplex on some vendors' devices, you are forced to also configure the speed, which shuts off Auto-Negotiation for both speed *and* duplex. On other vendors' devices, it's possible to configure the duplex setting while leaving the speed alone, allowing Auto-Negotiation to remain enabled for speed but not for duplex. The best advice is not to make any assumptions about how Auto-Negotiation works based on what you've seen from any one vendor. It's up to you to read the manual for each device and to understand what the various management commands will do.

It's also a good idea to avoid buying equipment from vendors with poorly designed Auto-Negotiation systems. Bad Auto-Negotiation implementations can be found on equipment from the largest vendors, so don't assume that buying switches from a major vendor means that you can ignore these issues.

Auto-Negotiation Debugging

The Auto-Negotiation protocol described in the Ethernet standard will, in the vast majority of cases, result in a correctly configured link. However, the confusion caused by misunderstanding the operation of Auto-Negotiation and Parallel Detection has led some people to believe that the Auto-Negotiation system is failure-prone. Further, the

problems encountered with poorly designed Auto-Negotiation implementations have caused some to conclude that Auto-Negotiation cannot be trusted to work correctly.

Once you understand the operation of Parallel Detection in combination with some vendors' implementations of Auto-Negotiation, then the occurrence of duplex mismatch problems becomes much less mysterious. And while it's true that there have been some buggy Auto-Negotiation implementations, equipment based on the standard has been shipping since 1995, which means that the vendors have had a long time to fix the bugs and to ship fixed interface drivers and switch software.

Nonetheless, you may still encounter implementations that have not been upgraded. One vendor, Cisco Systems, Inc., has a publicly available document for troubleshooting NIC compatibility issues with its Catalyst switches.[7] This document provides a listing of the major compatibility problems that Cisco customers have encountered with Ethernet interfaces from a variety of vendors.

While the list of problems in that document is fairly long, bear in mind that these incompatibilities have all been resolved, and that updated NIC drivers and Cisco software have been provided. The Cisco document notes that in addition to problems due to buggy software drivers or implementations, there have also been problems caused by some vendor-specific optional features, such as automatic signal polarity correction on Ethernet cables.

As things stand today, the vast majority of Auto-Negotiation implementation bugs that have been encountered in the field appear to have been fixed. Incompatibilities due to vendor-specific options can usually be resolved by shutting off all options other than Auto-Negotiation.

The majority of the remaining issues are most often problem reports of "broken" Auto-Negotiation that in actuality are reports of the duplex mismatch performance issues that arise with vendor implementations that shut off Auto-Negotiation due to manual configuration on one end of a link, while Auto-Negotiation is still enabled on the other end.

General Debugging Information

As of this writing, consumer-grade computers work quite well when it comes to Auto-Negotiation, and Auto-Negotiation problems on these devices are rare.

On the other hand, you may still encounter problems with Auto-Negotiation on some high-end servers or other equipment that is not sold in high volumes. Such equipment is typically used in more restricted environments. We can speculate that this leads to fewer encounters with other types of machines, and therefore to fewer bug reports and bug fixes for the high-end equipment.

7. Cisco Systems, Inc., "Troubleshooting Cisco Catalyst Switches to NIC Compatibility Issues," October 2009.

Another possibility is that, due to its initial expense and performance characteristics, this kind of equipment may be upgraded less frequently. This would help explain why older and buggier hardware and software seems to persist for considerably longer on these high-end servers than it does in the higher-volume consumer-grade equipment.

Media converters and Auto-Negotiation

When troubleshooting Auto-Negotiation issues on a link, be aware that if media converters are used to convert signals between twisted-pair and fiber optic segments on the link, you may find that Auto-Negotiation capability is included on the twisted-pair port of the converter. This can complicate troubleshooting, because it may not be obvious that there is a media converter located somewhere in the link.

When using media converters, you may wish to enforce a policy of manual configuration on the devices at both ends of the link, with Auto-Negotiation disabled on the media converter. This avoids the complexity of troubleshooting a link with an auto-negotiating device embedded somewhere in it. In the next section, we discuss how to develop policies to help ensure that Ethernet links are correctly configured at your site.

Debugging Tools and Commands

To discover the source of any configuration problem, you may need to use a variety of approaches and tools. Your goal is to find out whether a configuration problem is due to manual misconfiguration, missing Auto-Negotiation support on one end of a link, or something else. Here are some suggested approaches and tools that can be used:

Investigate log files
: A duplex mismatch will usually generate late collision errors, which may be logged as a counter on the Ethernet interface or switch port, or in an error log on an Ethernet switch. Examination of these logs for late collision errors will help determine if there are any problems, and if so, on which port(s). Many switches support remote logging of errors to a central computer, providing a single log file that can be used to check the error reports across a set of switches.

Use special management protocols
: One major vendor, Cisco Systems, Inc., has its own link layer management protocol that can help detect duplex mismatches. The Cisco Discovery Protocol (CDP) sends packets that contain information about the port configuration and capabilities of each switching port. This provides enough information about the port settings on each device connected to a link to allow the devices to detect when a duplex mismatch has occurred, and to send alerts and logging messages.

Run throughput testing software
: A network throughput tester is designed to send a rapid stream of packets through the link to test the maximum achievable throughput, which will usually reveal the

effects of duplex mismatch, as well as certifying link performance for correctly configured links. There are a number of good throughput testing programs; one of our favorites is *iperf*, which runs on Unix, Macintosh, and Windows computers.

Testing with a network throughput tool is essential, because normal web traffic may use so little bandwidth that the performance impairment is not immediately obvious. More performance analysis techniques can be found in Joseph D. Sloan's *Network Troubleshooting Tools* (O'Reilly, 2001).

Check the management interface or run a management program
: Often the quickest and simplest way to check for link misconfiguration is to log into an Ethernet switch and investigate the Auto-Negotiation and duplex settings, by looking at the switch management displays (e.g., "show interface," "show port"). A port or interface display may use the words "a-10" and "a-half" to indicate that these settings were configured by the Auto-Negotiation system, or it may show the manually forced speed and duplex settings when Auto-Negotiation is not enabled.

  It can be more difficult to investigate the configuration settings on the NICs or embedded Ethernet interfaces found in user computers and server machines. However, a number of vendors and operating systems do provide diagnostic and management programs for their Ethernet interfaces that allow you to display the configuration settings on the interfaces. That way, you can discover whether the interface is configured for Auto-Negotiation, and what settings have been automatically negotiated or manually configured.

  Finding the Ethernet interface settings on Microsoft Windows systems can be challenging. If you have the diagnostic software installed for the network interface, then running that software may display the interface configuration settings. If you don't have any interface diagnostic software installed, then you may be able to find a copy of the software on the website of the company that made the network interface in your computer.

  On Linux systems, the *mii-tool* and *ethtool* programs allow you to investigate the configuration of an Ethernet interface.

Troubleshooting Auto-Negotiation

Here are a few things you can try in order to troubleshoot Auto-Negotiation on a NIC or switch port:

Toggle Auto-Negotiation
: Disconnecting and reconnecting the patch cable, or using the management interface to disable and re-enable Auto-Negotiation on the NIC or switch port, will toggle Auto-Negotiation. By causing the Auto-Negotiation process to repeat, you can ver-

ify the resulting configuration of the NIC or switch port to see if there are Auto-Negotiation problems.

Try a different network interface
: On computers with removable NICs, try replacing the NIC with one from another vendor. This can help isolate the problem to a given NIC, indicating that the NIC driver software is buggy.

Upgrade the NIC driver software and the switch software
: Sometimes the easiest approach to debugging suspected Auto-Negotiation problems on a NIC is to download and install new NIC driver software from the vendor, then reboot the computer. Similarly, downloading and installing the latest stable release of software for an Ethernet switch may be required, especially if there are problems that appear on multiple switch ports.

Developing a Link Configuration Policy

The challenge for a network manager is to come up with a link configuration policy that results in stable, reliable, and high-performance networking at your site. To that end, a network manager needs to understand how Auto-Configuration works, and know how to avoid the most common mistakes when configuring Ethernet links. The following list summarizes what we've learned so far:

- Auto-Negotiation will result in the correct configuration of the highest-performance settings between devices on an Ethernet link, as long as Auto-Negotiation is enabled for all capabilities on the devices at each end of the link.

  Ensuring that Auto-Negotiation is enabled also helps ensure that MDI-X automatic signal crossover will continue to function when supported on twisted-pair interfaces.

- Manually configuring the speed and/or duplex may cause Auto-Negotiation to be disabled on the device, given the Auto-Negotiation implementations on some vendors' equipment.

  If a manually configured device with Auto-Negotiation disabled is connected to an auto-negotiating link partner, then the auto-negotiating link partner will default to half-duplex mode, because there are no Auto-Negotiation signals being received from the manually configured device at the other end of the link. Stated differently, if Auto-Negotiation is enabled only at one end of a link, the auto-negotiating device will always default to the half-duplex mode of operation. Because of this default behavior, a device without Auto-Negotiation support enabled that is set to full-duplex mode will always result in a duplex mismatch when connected to a link with an auto-negotiating device at the other end.

- Full-duplex mode, which provides maximum performance over the link, will automatically result if both devices on a link are configured to use Auto-Negotiation.

Link Configuration Policies for Enterprise Networks

Most sites have found that the best way to avoid any problems or performance issues due to duplex mismatch on Ethernet devices is to make sure that Auto-Negotiation is enabled on every device that supports it. If Auto-Negotiation is enabled on the devices at both ends of a link segment, then a duplex mismatch will not occur.

Similarly, if Auto-Negotiation is enabled on all desktop devices connected to Ethernet switch ports that also have Auto-Negotiation enabled, then there will not be a duplex mismatch on any of the desktop links connected to the switch. Note that this policy also assumes that all twisted-pair cabling in use is Category 5/5e or better, as is required to support 10, 100, and 1000 Mb/s speeds.

To avoid any issues with software, ensure that you have bug-fixed code by upgrading the software in your switches to the latest stable release, no matter which vendor you use.

Enabling the Auto-Negotiation protocol at both ends of the link ensures that the highest-performing settings common to both devices are automatically chosen and that both speed and duplex are set correctly, avoiding any problems with incorrect mode selection. By enabling Auto-Negotiation and avoiding manual configuration of speed or duplex settings, you will provide the best opportunity for correct link operation to occur.

Issues with Manual Configuration

The more ports that you manually configure, the harder it will be to ensure that the configuration will always be correct on both ends of the link, and the harder it will be to keep track of all of the settings. If you manually configure full duplex on all switch ports without also configuring all connected devices to match, then you can end up with duplex mismatches.

If you manually configure 100 (or worse, 1,000) switch ports, for example, then you also have to manually configure the 100 (or 1,000) devices connected to those ports. Further, you have to continually ensure the correct configuration of all devices that are ever changed, upgraded, or added to your network.

Given the difficulty of ensuring that everything anyone ever connects to an Ethernet port is going to be manually configured in the correct mode, the easiest and most reliable thing to do is to let Auto-Negotiation handle the task. Also, as the switch ports and desktop computers in a building are upgraded, as long as they have Auto-Negotiation enabled, they will automatically choose the highest-performance mode of operation without any need for manual intervention.

CHAPTER 6

Power Over Ethernet

Power over Ethernet (PoE) is an optional standard that provides direct current (DC) electrical power over Ethernet twisted-pair cabling. This makes it possible for an Ethernet switch port, for example, to provide both Ethernet data and the power needed to operate a low-power Ethernet device connected to the other end of the cable, such as a wireless access point. The system is carefully designed to provide both power and Ethernet data over the same cable, without causing any interference with the data.

PoE supports devices with relatively low power requirements, including wireless access points, Voice over IP (VoIP) telephones, video cameras, and monitoring devices, making it possible to reduce costs by avoiding the need to provide a separate electrical circuit for the connected device. The power being provided is classed as Safety Extra Low Voltage (SELV), which is defined as a voltage that is limited to a peak of 60 volts DC, provided by a power supply that has no direct connection to primary power (AC grid), and which derives its power via a transformer or equivalent isolation device.

In other words, the power being provided over the Ethernet cable is carefully engineered so as not to present a shock hazard. The low voltages, electrical isolation from the AC grid, and limited current levels mean that the power being delivered is safe to work with, and that you do not need an electrician to install or manage these power circuits.

Power Over Ethernet Standards

Power over Ethernet was first developed in 2003, in the 802.3af supplement, which became Clause 33 of the 802.3 standard.[1] The original 802.3af version of PoE is the most widely deployed version of the standard, and provides up to 15.4 watts of DC power for

1. The formal title of IEEE 802.3 Clause 33 is "Data Terminal Equipment (DTE) Power via Media Dependent Interface (MDI)."

transmission over the Ethernet cable. Power over Ethernet is defined to work over 10BASE-T, 100BASE-T, and 1000BASE-T links.

The current PoE standard does not specifically include (or exclude) 10GBASE-T links. However, a "call for interest" was issued in March 2013 to propose an extension of the PoE standard that will provide power over all four cable pairs simultaneously, and that will also be specified to meet the requirements of 10GBASE-T links. Based on the amount of work that needs to be done and the frequency of IEEE meetings, and assuming that work on the new specifications proceeds as expected, it is anticipated that the new standard may be completed by the spring of 2016.

A revision of the PoE standard was developed in 2009 in the 802.3at supplement, which extends the Clause 33 specifications to provide up to 34.20 watts. The new standard is also referred to as "PoE Plus," or "PoE+," and these nicknames may be seen in vendor marketing and documentation.

Vendors have developed their own extensions to the standards as well, to provide even more power. These include Cisco's "Universal Power over Ethernet," which provides up to 60 watts, as well as the "HDBaseT" specifications developed by an alliance of consumer electronics vendors, which includes a "Power over HDBASE-T" component that can deliver up to 100 watts over four pairs of Category 5e or 6 cabling. We will describe these vendor-developed extensions later in this chapter.

Goals of the PoE Standard

The goals listed in the IEEE PoE standard include:

Power
: Provide both power and data through the twisted-pair cable.

Safety
: Ensure that only safe (SELV) power is allowed onto the cable, isolating the cable from any other power source.

Compatibility
: Work without modification over existing twisted-pair Ethernet systems.

Simplicity
: Add no complexity for the end user beyond what is needed to connect a twisted-pair Ethernet link.

Devices That May Be Powered Over Ethernet

Many access points, telephones, and video cameras can be powered over the original 802.3af PoE system that provides roughly 15 watts (the powered device may draw a maximum of 12.95 watts to allow for power losses over the cable). However, access points supporting the newer 802.11 standards and containing multiple radios, or video cameras that have motors for zoom, pan, and tilt functions, can draw more than 12.95 watts. The 802.3at revision of the PoE standard provides up to 34.20 watts, resulting in a minimum of 25.50 watts at the Powered Device.

Meanwhile, the widespread success of PoE has resulted in demands to support devices that have even higher power requirements, such as display monitors, medical monitoring devices, and building automation systems (door locks, card-key systems, HVAC monitoring). The adoption of light-emitting diodes for display monitors and general lighting has reduced power requirements for those devices, increasing the number of devices that could be powered with PoE if more wattage could be provided.

As mentioned previously, to meet these increased demands, several vendors are already providing vendor-specific versions of four-pair PoE that provide high power levels (up to 60 watts, or even higher). These systems may not be interoperable. When the new IEEE standard is completed, it will provide interoperable and vendor-neutral technology for achieving higher power over four cable pairs.

Benefits of PoE

PoE capability saves costs by avoiding the need for installation of power circuits for things like wireless access points. Providing dedicated power circuits to hundreds or even thousands of APs located in the ceilings of buildings across a campus network system would be very expensive.

Power over Ethernet has been instrumental in the adoption of 802.11 wireless LANs, by making it possible to provide power and data over a single cable. This vastly simplifies installation and reduces the cost of an access point installation. With PoE, you can provide power anywhere there is an Ethernet cable, providing far more flexibility than would hard-wired electrical outlets.

PoE also improves remote management, monitoring, and troubleshooting. PoE switches with management interfaces make it possible to provide manual power management of the powered devices. Using the switch management features, you can find out from the switch whether or not a Powered Device is drawing power, and how much power it is using. You can also control power to a device such as an access point by sending management commands to the switch, making it possible to toggle power on and off as part of troubleshooting efforts.

PoE Device Roles

The PoE standard describes two device roles:

Power Sourcing Equipment (PSE)
: The PSE is the device that provides power over a twisted-pair Ethernet cable. A PSE may be an Ethernet switch port, or it can be an outboard power injector.

Powered Device (PD)
: The device being powered by the PSE. A Powered Device may include a wireless access point, VoIP telephone, or IP-based security camera. The Powered Device is also called Data Terminal Equipment (DTE) in the standard.DTE

Standard Power Sourcing Equipment provides roughly 48 volts of direct current power to the Powered Device through two pairs of twisted-pair cabling. Figure 6-1 shows the two basic methods of sourcing power over Ethernet: endpoint and midspan. With an endpoint PoE device, the Power Sourcing Equipment is also the termination point of the Ethernet link, such as a PoE switch port.

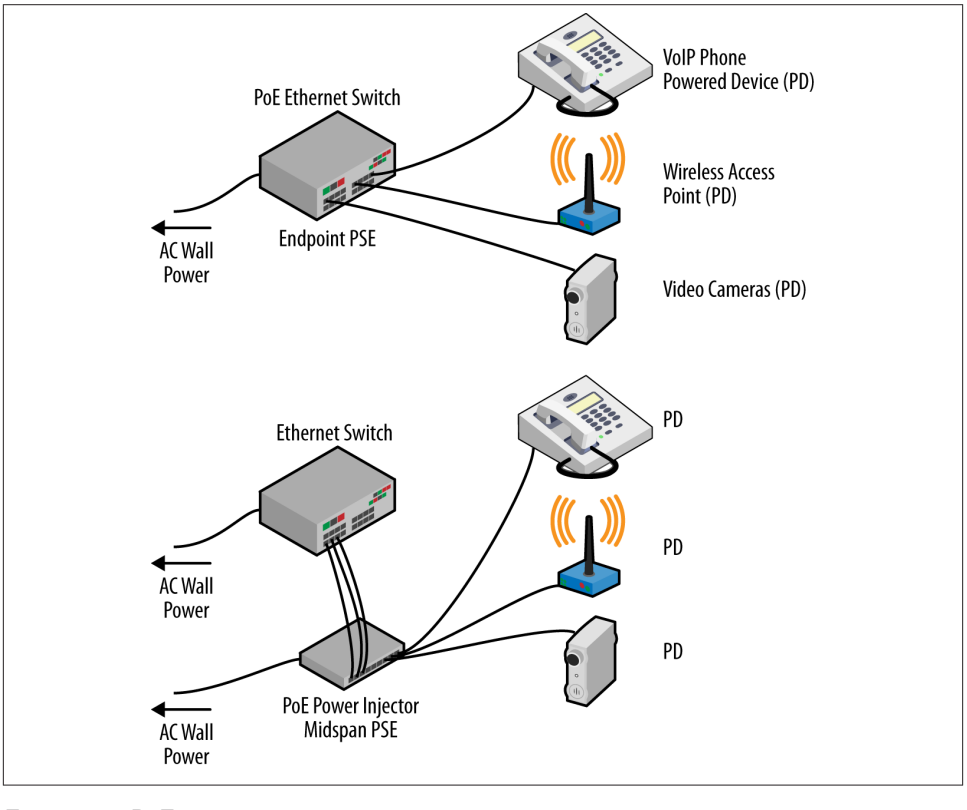


Figure 6-1. PoE connections

If the switch does not provide power over Ethernet, then an outboard power injector, called a *midspan PSE*, can be used. Despite the name, the power injector does not have to be located in the middle of the link. Instead, it can be located anywhere along the link that is both convenient and provides an AC power connection, which the midspan PSE converts to DC power for injection into the Ethernet link.

Midspan power injectors typically come in two forms: single port and multiport. Multiport midspan power injectors provide power to multiple Ethernet devices. This device appears as a box that plugs into AC wall power and converts the wall power into DC power for injection into the Ethernet link. You can also use a single-port midspan injector, which plugs into AC wall power and provides DC power to a single device over an Ethernet link, such as a wireless Ethernet access point.

Figure 6-2 shows a PoE *splitter*. This is a Powered Device that connects to a PoE port and splits out the DC power as a separate connection that is typically based on a round electrical plug that connects to an external device. This makes it possible to provide DC power and Ethernet data over the same cable to a non-PoE device that has an Ethernet connection and a standard DC power plug connection.

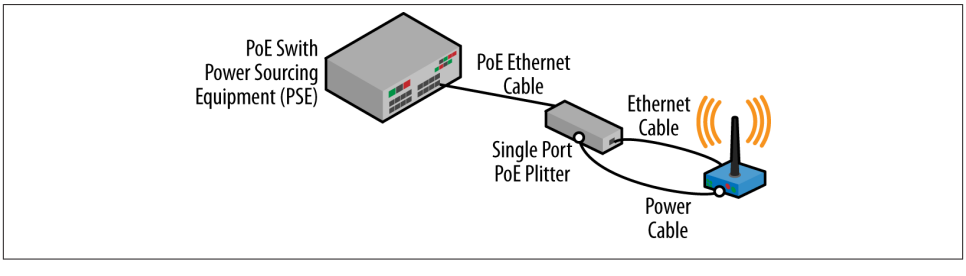


Figure 6-2. PoE splitter

PoE Type Parameters

The original PoE standard was defined to provide 12.95 watts at the Powered Device over 100 meters of Category 3 or Category 5 cable. The 802.3at extension provides 25.50 watts at the PD over Category 5 or better cabling. Category 3 cabling is not supported for 802.3at PoE.

To help organize the different specifications, the standard defines two types of Power over Ethernet systems, called Type 1 and Type 2, with different parameters for each. Type 1 refers to the original, low-power system, and Type 2 refers to the newer system that provides higher power.

Table 6-1 shows the key parameters involved in Type 1 and Type 2 PoE power systems. A "power system" is defined in the standard as a single PSE, link segment, and PD connected together. While Type 1 operation works over old Category 3 cabling and adds

no significant requirements to the cabling, Type 2 operation requires Category 5 or better cabling and a derating of the maximum ambient operating temperature for the cabling system.

Table 6-1. PoE Type 1 and Type 2 systems

Property	Type 1	Type 2
Power at the PD[a]	12.95 W	25.50 W
Maximum power provided by PSE	15.40 W	34.20 W
Voltage range at PSE	44.0–57.0	50.0–57.0
Voltage range at PD	37.0–57.0	42.5–57.0
Maximum current	350 mA	600 mA
Maximum cable loop resistance	20 ohm (Category 3/Category 5 or better)	12.5 ohm (Category 5 or better)

[a] The power supply at the Powered Device will consume a small percentage of the available power (typically 10%), leaving the remainder for the PD.

PoE Operation

The PoE standard defines a method for powering a PD over a cable by Power Sourcing Equipment, and then removing power when the PD is disconnected. The formal process includes an idle state and three operational states: *detection*, *classification*, and *operation*. The PSE leaves the cable unpowered (idle state) while it periodically checks to determine if a PD has been connected to the cable. This process is called detection.

Once a PD has been detected, the PSE may execute a probing process, called classification, to determine the current levels needed by the PD. If the PSE has enough power, then it begins the operation phase by powering the PD. It then monitors the power levels to ensure that the PD is still connected.

Power Detection

There are several challenges when it comes to providing power over an Ethernet cable, with the first being: how do you know what's at the other end of the cable? If you just provide power without checking, you could damage any non-PoE equipment that might be attached to the end of the cable.

After all, there's no guarantee that the twisted-pair cable you are powering is connected to an Ethernet device. The cable might be connected to an analog telephone instead, which might be sensitive to power on its cable connection. If an installer tries one cable after another in an attempt to get something to work (which is not an uncommon approach), then a powered cable could cause a problem when connected to something that isn't an Ethernet device.

To avoid placing power on the link unless there is a Powered Device connected to the other end of that link, the standard provides a method, called *power detection*, for detecting the presence of a Powered Device. Once a Powered Device is detected, then another set of mechanisms, called *power classification*, can be used to determine what power level is needed by the PD.

Power detection is performed by the Power Sourcing Equipment, which periodically monitors the Ethernet link for the presence of a Powered Device at the other end. This is done by applying a small voltage (2.70 V to 10.1 V) to the cable pairs and measuring the current flow, looking for the presence of a 25,000 ohm resistance provided by the PD as a signal that it is present.

A low voltage is used for detection because it is unlikely to cause damage to devices not designed for PoE that might be connected to the link. If a 25,000 ohm resistance is detected on the cable pairs, then this is considered a valid PD signature, indicating the presence of a Powered Device at the other end of the link. The next thing that the PSE does is to determine how much power to send over the link.

Power Classification

After the PD is detected, the PSE and PD can interact to determine how much power the PD requires. There are two power classification mechanisms that are used: physical layer classification and data link layer classification. Type 2 PDs that require more than 13.0 watts must support data link layer classification, which is optional for all other devices. If both classification systems are supported by the PSE and the PD, then the information provided by data link layer classification takes precedence over that from the physical layer classification.

If a PSE has multiple ports, then the typical approach is to probe one port at a time, classify the power required, ramp up the power to the required amount, and move on to the next port until all the power that can be provided by the PSE is allocated. This makes it possible to bring up a port at the full power level supported for a Type 2 PD, for example, and then negotiate the actual power level needed by the PD using the data link layer classification system.

Physical layer classification method

The physical layer classification was defined in the original 802.3af system, and all Type 1 PSEs may optionally use the physical layer classification system to classify the power requirements of the PD. If the Type 1 PSE doesn't support classification, then it simply provides the full power level of 15.4 watts whenever it detects a Powered Device. While that provides the simplest method of automatic PoE operation, most vendors use PSE controller chips that provide a power management capability to classify the power requirements on a link.

The physical layer classification process occurs before a PSE provides power to the PD over the link. The process consists of the PSE applying a reduced voltage (between 15.5 V and 20.5 V) into the cable pairs, and measuring which one of a limited set of currents is being drawn by the Powered Device. The specific current detected is used as a signal provided by the PD to inform the PSE as to the current requirements of the PD.

This makes it possible for the PSE to determine which of several current levels is required by the Powered Device before actually providing the full voltage and power levels to the link. After the classification process completes, there is a "ramp up" process that increases the power to the level determined by the classification process.

Table 6-2 shows the five power classifications defined in the standard. If the Type 1 PSE does not support classification, then it must assign all PDs to Class 0, which is 15.4 watts.

Table 6-2. Physical layer power classifications

Class	Usage	Current (mA)	Minimum watts at output of PSE	Watts at PD	Description
0	Default	0–4	15.4	0.44–12.95	Classification unimplemented
1	Optional	9–12	4.00	0.44–3.84	Very low power
2	Optional	17–20	7.00	3.84–6.49	Low power
3	Optional	26–30	15.4	6.49–12.95	Mid power
4	Type 2 devices	36–44	36.0	12.95–25.50	High power

A Type 2 PD presents a signature that indicates that it is a Class 4 device. A Type 1 PSE will treat a Type 2 PD as a Class 0 device, and provide 15.4 watts if it has the power available. A Type 2 PSE will interact with a Type 2 PD such that both ends detect a Class 4 device; the PD will know that it is connected to a high-power PSE and may draw up to 25.5 watts. A Type 2 PD that does not receive the Class 4 Physical Layer signatures may choose not to start, or must start at a power level of 13 watts and request more power via the data link layer classification system after startup.

Data link layer classification method

The 802.3at extension to the standard defines a separate classification mechanism called data link layer classification that is based on the Link Layer Discovery Protocol (LLDP), transmitting LLDP packets with "Organizationally Specific TLVs" (where TLV stands for type-length-value). These packets are defined in Clause 79 of the 802.3 standard.

The LLDP packets allow the network management functions supported by the PSE and PD to advertise and discover the power requirements of the Powered Device. The 802.3at extension defines LLDP information packets that can advertise this information between the PD and the PSE.

Type 2 devices are required to do both Type 1 physical layer classification and the data link layer classification. The Type 2 device will use the physical layer classification system

to identify itself as a Class 4 device, requiring high power. The Type 2 Powered Device must not draw more than 13 watts if it has not received a Type 2 Class 4 Physical Layer signal.

Type 2 powered devices that can operate with less than 12.95 watts must be able to be powered by a Type 1 PSE. If a Type 2 PD cannot operate at less than 12.95 watts, when connected to a Type 1 PSE it must indicate to the user that there is not sufficient power available over the link for the device to function correctly.

The LLDP packet exchanges also provide an optional capability to dynamically adjust power requirements after the initial detection and classification phases. Dynamically adjusting power can help manage the total amount of power that must be provided by a PoE switch on its Ethernet ports to meet the needs of the PDs, making the system more efficient. Note that the dynamic power changes are not designed for rapid variations; a PSE may take as much as 10 seconds to respond to a requested power change.

Mutual identification

The interrogation and power classification functions make it possible for the PSE and PD to provide "mutual identification," in which each determines whether it is connected to a Type 1 or Type 2 device. Mutual identification also provides useful information for devices that support power management features instead of purely automatic PoE.

If a PD or PSE does not implement classification, it will not be able to complete a mutual identification process and will only be able to perform as a Type 1 device.

Link Power Maintenance

Once the detection and classification processes have completed, then the PoE link is operational and the PSE is providing power over the link. The PD provides a *maintain power signature* (MPS) consisting of both the current draw used by the PD and a specified input impedance, which can be sensed on the cable pairs by the PSE.

This makes it possible for the PSE to monitor the link for the continued presence of an active PD. If loss of MPS is detected, then the PSE will quickly remove power from the link and return to the detection process of monitoring the link for the presence of a PD.

Power Fault Monitoring

While providing power, the PSE also monitors the link for various fault conditions, including under- and over-voltage conditions and under- and overcurrent conditions. When any of these conditions is detected, the PSE will shut off DC power to the link and go back to the power detection phase.

Power shutoff is designed to happen rapidly, in a maximum of one half second. This helps to avoid supplying power to a cable that has been disconnected and is then quickly reconnected to a different port in the cabling system.

PoE and Cable Pairs

The PoE standard specifies the way that power is injected into the Ethernet cable pairs, and which pairs should be used. Power is injected using one pair of wires to carry the positive current ("plus"), and a second pair to carry the negative ("minus"). This provides two wires to carry the positive power and two wires for the negative power, which helps reduce resistance and heating effects. The standard defines two alternatives for which cable pairs are used.

Alternative A uses the data signal pairs (wires connected to pins 1,2 and 3,6). Power is placed on the data pairs by connecting the DC power supply to a center tap on the internal signal coupling transformers, an approach that is also called "phantom power."

Alternative B uses the "spare" pairs (wires connected to pins 4,5 and 7,8). The spare pairs are so called because they do not carry data signals in the 10BASE-T and 100BASE-T systems. When the spare pairs are used, the power is coupled directly to the wire pairs, with no need for a transformer. Note that all four pairs are used to carry signals in the 1000BASE-T systems, meaning that transformer coupling is used on all four wire pairs, and that the only difference between Alternatives A and B in 1000BASE-T systems is the choice of cable pairs that are used.

Some vendors have developed their own four-pair systems of Power over Ethernet. These products are based on vendor-developed "standards," and typically require that the equipment at both ends of a PoE Ethernet link come from the same vendor or group of vendors. A new IEEE standard that is currently under development will provide a vendor-independent four-pair system of PoE that will interoperate with any vendor's gear that follows the IEEE standard.

For a power sourcing device, under the current PoE standard either alternative can be used, but not both. The Powered Device must be able to accept power on either the Alternative A or the Alternative B wiring scheme, as the PD cannot know which pairs the PSE may use.

Also, the PD must be able to accept either current polarity (positive or negative) on a given wire pair. That's because there is no way to guarantee which polarity of power will end up on which PD wire pairs, given that the switch port may implement automatic signal crossover (MDI-X), or that a patch cable wired to provide a crossover may be present in any given Ethernet link. Therefore, the PD contains circuitry that automatically accommodates all possible wiring variations with respect to which pairs are car-

rying DC power, as well as all possible variations with respect to the polarity of the current carried by any given pair.

As shown in Figure 6-3, the 10BASE-T and 100BASE-T standards use only two pairs, 1,2 and 3,6.

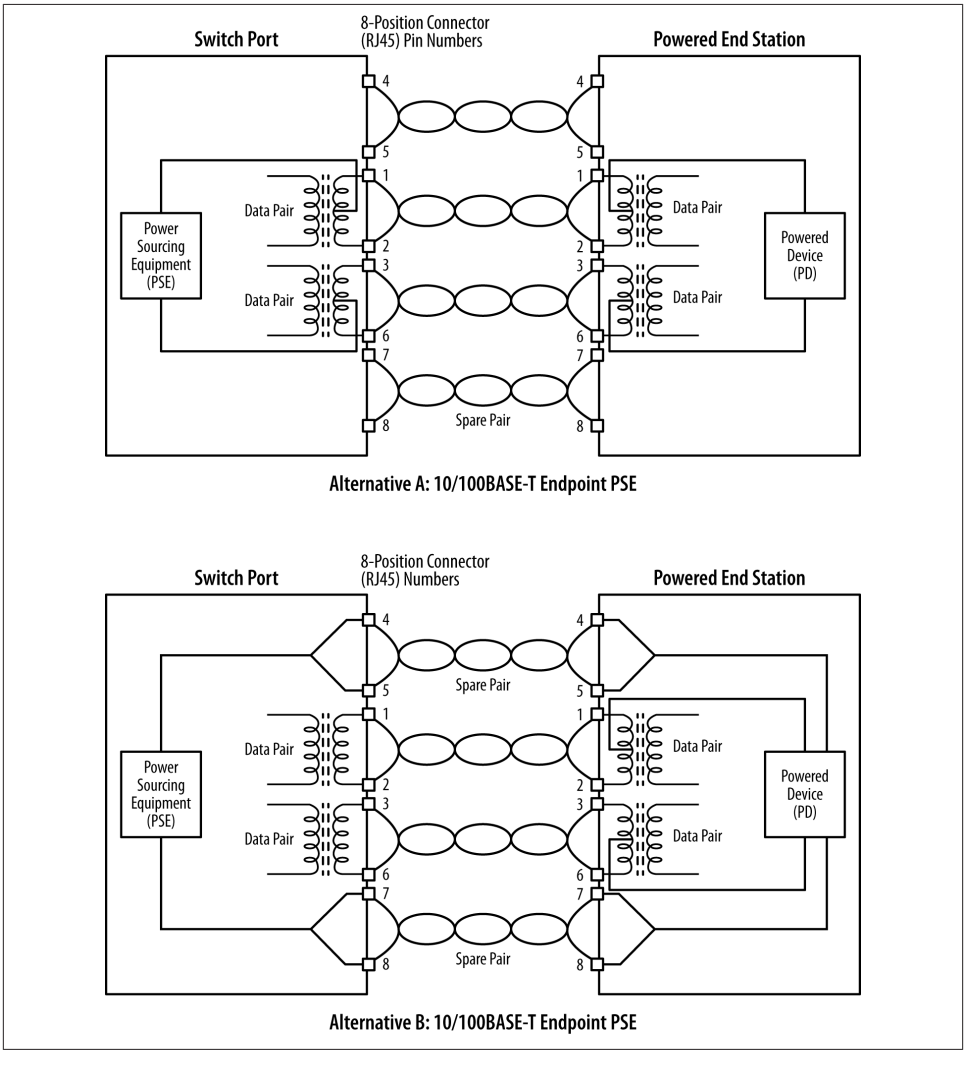


Figure 6-3. Alternatives A and B for 10BASE-T and 100BASE-T

Figure 6-4 shows that the 1000BASE-T standard uses all four pairs.

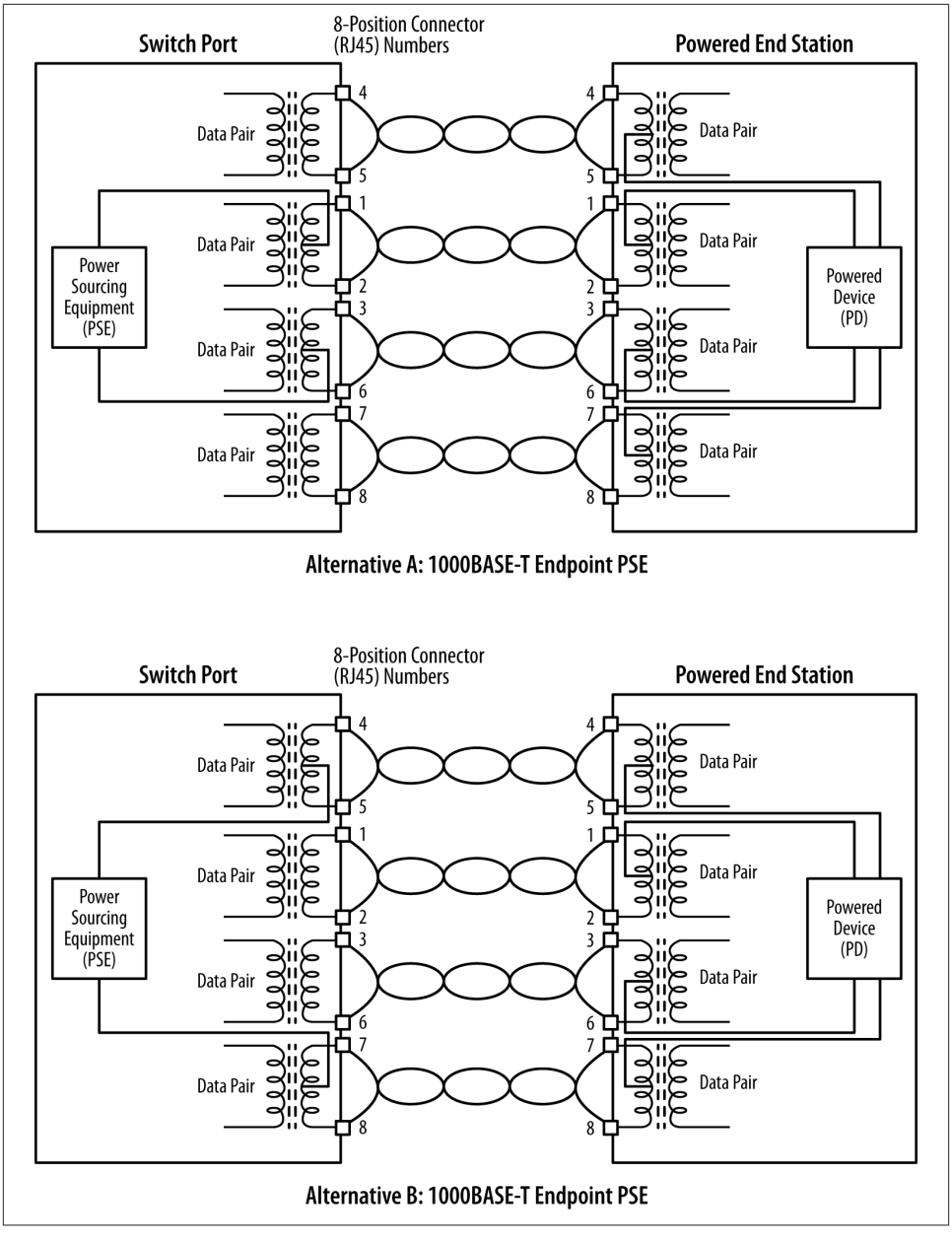


Figure 6-4. Alternatives A and B for 1000BASE-T

The wiring schemes used for midspan devices look much the same, with the major difference being that the midspan device is inserted in the cabling path between the Ethernet switch and the Powered Device. The midspan device contains the DC power

supply and injects the power into the Ethernet link using the same Alternative A and B wiring schemes as for 10/100BASE-T and 1000BASE-T links.

PoE and Ethernet Cabling

The original 802.3af PoE standard was designed to operate over Category 3 cabling, which was defined as having two pairs available for use by Ethernet signals. The 802.3at extension to the PoE standard specifies Category 5e cabling or better. This is more formally referenced as Class D or better cabling, as specified in the international ISO/IEC 11801:1995 standard.

One technical issue with Power over Ethernet is that it causes a small amount of heating of the cable to occur, due to the resistance of the copper wires interacting with the DC power being carried over the wires. The slightly elevated temperatures do not present any safety risk, or risk of damage or premature aging of the cables or connectors.

When many cables are bundled together and all cables are carrying maximum amounts of PoE power, however, the small amount of heating becomes more significant in that it could affect the signal-carrying characteristics of the cables in the bundle. Along these lines, it's important to note that Category 6 and 6A cabling both have thicker copper wires than Category 5, and that thicker wire results in lower resistance to DC currents and therefore significantly less heating. Some tests indicate that the amount of heating in Cat6/6A cables is approximately 50% less than in Cat5/5e cables. By the way, tests also show that there is no problem using shielded twisted-pair cables to carry PoE, and that the foil or stranded shielding actually helps the cable to dissipate any heating effects.

Modifying cabling specifications

The development of the 802.3at standard included careful consideration of what happens in large cable bundles with PoE enabled on all cables. However, the cabling standards are not part of the IEEE standards, and cabling specifications are not the responsibility of the IEEE. The IEEE standard does note that:

> *Under worst-case conditions, Type 2 operation requires a 10°C reduction in the maximum ambient operating temperature of the cable when all cable pairs are energized, or a 5°C reduction in the maximum ambient operating temperature of the cable when half of the cable pairs are energized. Additional cable ambient operating temperature guidelines for Type 2 operation are provided in ISO/IEC TR29125 and TIA TSB-184.*[2]

As a result of these considerations, one recommendation is to avoid high temperatures around the cable bundles, and to ensure that the operating temperature around cable bundles does not exceed 50°C (122°F).

2. IEEE Std 802.3-2012, paragraph 33.1.4.1, p. 622.

The cabling considerations for PoE are described in separate standards documents. The two documents cited by the IEEE include a technical bulletin entitled "TSB-184 Guidelines for Supporting Power Delivery Over Balanced Twisted-Pair Cabling," which is available from the Telecommunications Industry Association, and an ISO standards document called ISO/IEC TR 29125, "Information technology—Telecommunications cabling requirements for remote powering of terminal equipment," which is available from the ISO. Both documents provide guidance to cable installers to help characterize the amount of heating that can be expected in the worst case (large bundles with lots of cables carrying power) and to ensure that a cabling system will work optimally when PoE is in use.

The bottom line is that a small amount of heating occurs on any copper cable that carries power. The heating is not an issue for individual Ethernet cables, but when those cables are bundled tightly together in a cabling system, and if all cables are carrying the maximum amounts of power, then the increase in temperature could have an effect on signal quality, especially if the ambient temperature around the cables is already very high. Under normal circumstances with cables inside a typical office building, the cables will function correctly.

PoE Power Management

An Ethernet switch that provides PoE over some or all of its ports can also provide a power management point for the network devices powered over the Ethernet links. While the PoE standard defines the mechanisms for sending direct current power over Ethernet cables, it does not define which options and other management capabilities vendors may implement in the equipment that they build. Nor does it define how those management functions may be designed, or what the management commands may look like. Assuming that a management interface is provided, the documentation for your switch or midspan equipment is your source for the details on how the interface is organized and what commands it uses.

As the PoE standard is defined to operate automatically, it's possible for a low-cost switch to provide a simple PoE capability that supports PoE operations but does not provide any management interface. However, many vendors do provide a management interface on the PSE that allows you to do such things as control which ports provide power, turn power on and off on a given port, and monitor power consumption per port.

PoE Power Requirements

As a user of PoE, it is important to understand that the power supply in the Ethernet switch or multi-port midspan device will be a limiting factor. In other words, you cannot provide more PoE power than the internal power supply of the PSE can handle. It's up to you to make sure that the PSE—typically an Ethernet PoE switch—has enough power to meet your needs.

For example, a 24-port switch providing 802.3af PoE at 15.4 watts per port requires an extra 370 watts of power beyond what is required to run the Ethernet switching functions in order to provide power to all 24 PoE ports simultaneously. If those ports each provide 30 watts of 802.3at power, then the switch will need 720 watts of extra power capability for the PoE ports.

PoE Port Management

To help manage these power loads, vendors of managed switches typically program the switches to provision power one port at a time. Allowing the switch to power up one port at a time avoids having to provide full power to all ports simultaneously, and potentially running out of power in the process.

As an example of how PoE switches are designed to operate, one major vendor of PoE switches provides three power management modes that can be configured on the switch ports:

Auto
: With this default setting, the switch automatically detects if the connected device requires power. If the switch has enough power, then it grants power to the port, updates the power budget information, and turns on port power on a first-come, first-served basis. If granting power to a given port would exceed the power budget, then the switch denies power, disables power on the port, generates a log message, and updates the port's LEDs to indicate the status.

Static
: With this setting, the switch preallocates power to a port even when no device is connected, and guarantees that power will be available for the port. A port with a static power configuration does not participate in the first-come, first-served model of operation. Because power is preallocated, any device that uses less than or equal to the preallocated amount is guaranteed to be powered when connected to a port configured as static. The statically configured power level is not allowed to be adjusted via power classification or messages carried by the CDP or LLDP protocols.[3]

Never
: This configuration disables PoE detection and turns the port into a data-only port.

PoE Monitoring and Power Policing

Vendors can also provide mechanisms for managing ports after they are powered. One major vendor's implementation includes power monitoring and policing. If a Powered

3. The Cisco Discovery Protocol (CDP) is Cisco's proprietary precursor to the LLDP protocol standard.

Device attempts to consume more power than the maximum allocated on a port, then the switch can be configured to turn off power on the port, or to just log the event and update the LED status light for the port.

The power policing feature is disabled by default. When this feature is enabled, it uses several pieces of information to determine what the power limit should be. You can configure a maximum wattage level for one or more individual ports (or for all ports on the switch), which becomes the power level at which policing will occur, or you can set an automatic or static power level on each port, which will allocate power dynamically or statically, as you prefer. Finally, you can let the switch port automatically determine the power usage of the Powered Device.

When power policing is configured, the switch polices the power usage sensed at the switch port, which is different than the amount of power that actually arrives at the Powered Device over the cable. The cable distance between the switch port and the Powered Device is responsible for some amount of power loss due to the resistance of the copper cable.

If a device uses more than the maximum power allocated on the port, the policing function on the switch can be configured to take different actions: it can be set to turn off power to the port, or it can be configured to just generate a log message and update the status on the port LEDs.

Power policing should not be confused with the overcurrent shutdown capability in the PoE standard. Power policing refers to configuring a power level on a port, and then having the switch take some action if that power level is exceeded. No matter what level you set, if a Powered Device consumes too much current and causes an overcurrent condition, then the PoE standard requires that the power be removed and that the port go back into power detection mode.

When policing is enabled, a PoE switch can also be configured to manage the order in which ports are powered. If automatic power policing is configured, then the switch will enable power one port at a time using ascending port numbers: port 1 will be powered first, then port 2, and so on.

It's possible for a switch to power off ports as well. If a new module is installed in a chassis switch, for example, and that switch now has less power to provide to its PoE ports, then typically the ports will be unpowered in descending order, starting with the highest port number that is powered and working downward until the power budget is sufficient to meet the power required by the remaining ports.

Vendor Extensions to the Standard

After 802.3at made it possible to provide roughly 30 watts over two pairs of wires, it was a relatively straightforward process to extend the standard to provide power over four pairs simultaneously. This approach provides roughly 60 watts of power, by implementing two sets of 802.3at-compliant electronics at each end of the link and powering all four wire pairs at the same time.

When the new IEEE four-pair PoE extension that is currently under development becomes adopted as part of the standard, it will become possible to provide higher power using technology that will interoperate between multiple vendors. Until that happens, various vendors have provided their own solutions, several of which are described next.

Cisco UPoE

Cisco Systems is one major vendor providing four-pair PoE power, which it terms Universal Power over Ethernet (UPoE). Cisco has also extended the LLDP-based power negotiation protocol to allow mutual identification and dynamic power budgeting up to 60 watts. Cisco switches can be configured to statically set the port power budget to enable support for devices that do not know how to deal with Cisco's UPoE extensions.

Microsemi EEPoE

Microsemi is a company that provides midspan PSEs for PoE. It has also extended the 802.3at standard to provide power over four pairs of cables, in an approach that it calls Energy Efficient Power over Ethernet (EEPoE). Microsemi points out that midspan PSEs can be upgraded without having to change your switching hardware, allowing an innovation like EEPoE to be implemented at lower cost (assuming you are already using midspan technology).

The advantages of this include more flexible power management than that provided by some PoE switches, as well as improved integrated circuits for PoE ports that minimize the power consumed by the PoE detection process, among other things. The vendor notes that four-pair powering helps reduce power losses over the copper cabling, further improving energy efficiency.

Power over HDBaseT (POH)

The HDBaseT specifications are not a part of the IEEE standard. Instead, they have been developed by the HDBaseT Alliance, which was formed to develop technology for use in home entertainment systems—hence the use of "high definition" (HD) in the specification name. The HDBaseT system operates over Category 5e/6 cabling to provide up to 10.2 Gb/s of uncompressed video and audio signals, 100BASE-T Ethernet, control signals, and power on the same cable.

The Power over HDBaseT (POH) portion of the system is based on 802.3at and uses all four pairs of cables to deliver up to 100 watts of power over distances up to 100 m, for powering home entertainment systems. The Energy Star specifications have been reducing the power required by television screens, with the most recent Energy Star 6.0 version limiting power consumption to under 100 watts for 60-inch screen sizes. This makes it possible to power many HD screens using the POH system provided by the specifications developed by the HDBaseT Alliance.

PART II
Ethernet Media Systems

This part will describe the Ethernet media systems. We begin with the basics of Ethernet media signaling in Chapter 7, where we also cover the Energy Efficient Ethernet system, which saves power by modifying the media signaling during idle periods.

Chapters 8 through 14 describe specific media systems. Each of the media system chapters covers the media components used for a given speed of Ethernet. The format of each chapter is identical, which helps to organize and clearly present the information. While an effort was made to avoid needless duplication of text, the format leads to some unavoidable repetition in these chapters. This is especially noticeable if you read several media chapters in a row.

CHAPTER 7

Ethernet Media Signaling and Energy Efficient Ethernet

This chapter introduces the media signaling components in the standard, and the Energy Efficient Ethernet extensions that modify the Ethernet signaling to save power when no data is being sent. Knowing how the media signaling components are organized and what they are called is helpful for understanding the ways Ethernet interfaces are connected to the various media systems and how they send signals over an Ethernet link.

To send Ethernet signals from one station to another, stations are connected over a cabling system based on a set of standard signaling components. Some of these are hardware components specific to each media cabling system. These media-specific components are described in more detail in the individual media chapters and cabling chapters that follow.

Other signaling components, such as Ethernet interface electronics, are common to all media systems. The standard refers to the specifications for these elements as "compatibility interfaces," because they ensure that stations can communicate in a compatible manner. The common signaling elements are described in this chapter.

Figure 7-1 shows a logical diagram of Ethernet stations A and B, connected over a link, with the physical layer standards involved shown in gray. The physical layer standards include further sublayers that are shown in more detail in Figure 7-2. Each station implements the same set of physical layer standards.

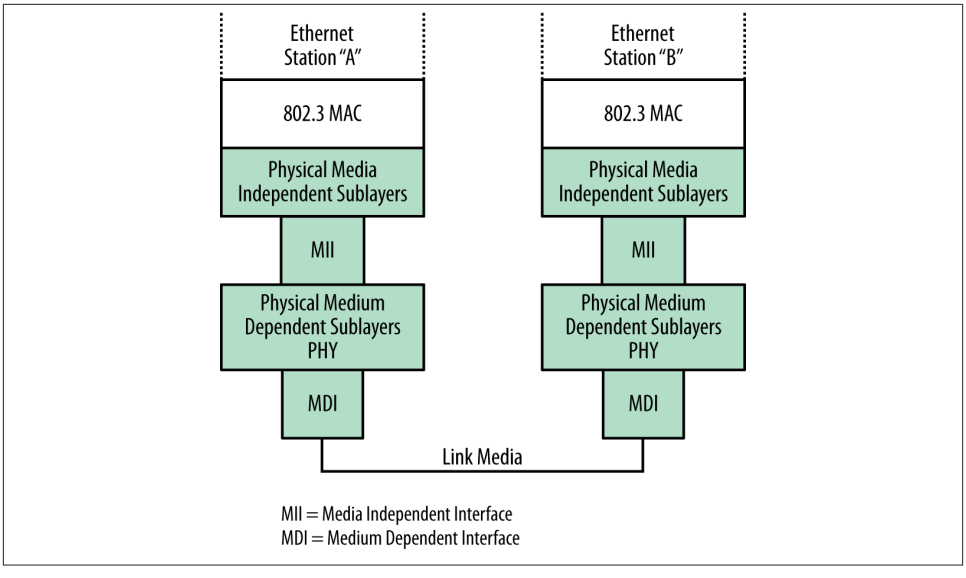


Figure 7-1. Ethernet physical layer standards

These sublayers are used for specifying the operations of the signaling and other mechanisms used to make the Ethernet link function for the given media speed being standardized. The sublayers help divide the task of sending signals over the physical media into further sections, some of which are independent of the media system and some of which depend on the specific medium involved.

The standard defines *medium dependent interfaces* (MDIs) for connections to each media system, and notes that all stations must rigidly adhere to the specifications for the physical media signals carried over the MDIs that are defined in the clauses of the standard that describe each Ethernet media system. The components used to couple signals directly onto the media are part of the physical medium dependent sublayers, also called the "PHY" (pronounced "fie").

As the Ethernet system has evolved, it has developed a set of *media independent interfaces* (MIIs) for each media speed. This means that this portion of the Ethernet interface is not specific to a cabling system. These interfaces are one of the ways the standard provides for a given Ethernet interface to be connected to different types of cabling.

For example, an Ethernet switch port (interface) could be equipped with a transceiver connecting it to a twisted-pair link or a different transceiver connecting it to a fiber optic link. Both transceivers are MDIs that connect to the same switch port (but not at the same time). The switch port electronics contain the MII, which interfaces to multiple MDIs. With this, transceivers for multiple media systems can be developed without requiring changes in the Ethernet interface electronics in the switch port.

Media Independent Interfaces

The first media interface developed for the 10 Mb/s Ethernet system was called the "transceiver cable" in the DIX standard, but its name was later changed to *attachment unit interface* (AUI) in the IEEE standard. The AUI supports the 10 Mb/s media systems only, connecting to all media types: coaxial cable, twisted-pair, and optical fiber.

The next media attachment standard was developed as part of the Fast Ethernet standard, and this was the first time the IEEE standard used the term "media independent interface." The 100 Mb/s MII provides support for both 10 and 100 Mb/s media segments. Both the AUI and MII standards included provisions for an external *medium attachment unit* (MAU), also known as a *transceiver*. The external MAU was connected between the cabling system and the Ethernet interface.

With the development of twisted-pair Ethernet that connects directly to RJ45 ports on Ethernet switches and other devices, external MAUs, or transceivers, connected to interfaces with external AUI cables (also called transceiver cables) are no longer used for Ethernet over copper cabling. A description of the older 10 and 100 Mb/s external transceivers can be found in Appendix C.

Next in the evolution of Ethernet, a *gigabit media independent interface* (GMII) was developed as part of the Gigabit Ethernet system. The GMII accommodates the increased speed of the Gigabit Ethernet system by providing an electrical definition for signals with a wider data path to the Ethernet interface, allowing them to carry more information as required by the faster speed. This set of signal paths is internal to the Ethernet interface and is not exposed to the user.

The MII data path has since been expanded to accommodate ever-faster versions of Ethernet, leading to multiple names for this element of the standard, which is now known as "xMII." Table 7-1 shows a set of xMII names, taken from the various physical layer standards.

Table 7-1. xMII versions

xMII version	Description
MII	100 Mb/s Media independent interface
GMII	1 Gb/s Media independent interface
XGMII	10 Gb/s Media independent interface
XLGMII	40 Gb/s Media independent interface
CGMII	100 Gb/s Media independent interface

Ethernet PHY Components

As the standard has evolved, more xMII versions have been developed, and more elements have been defined in the physical layer to provide more options for internal interconnections and to make the signaling function over faster media systems. The Ethernet interface chips can now support a whole series of xMII signaling interfaces, which are brought into play depending on which speed of Ethernet is chosen. This enables the Ethernet interface to implement a range of physical layer elements, depending on the media speeds that the vendor chooses to support.

As shown in Figure 7-2, the physical layer includes a physical coding sublayer, which specifies the signal encoding required for a given speed of Ethernet. It also includes physical medium attachment and physical medium dependent standards, which vary depending on the media type (copper or fiber optic). Also included as needed can be a forward error correction element to make signaling work better at higher speeds, and Auto-Negotiation, which may be required on some media types, optional on other media types, or unused.

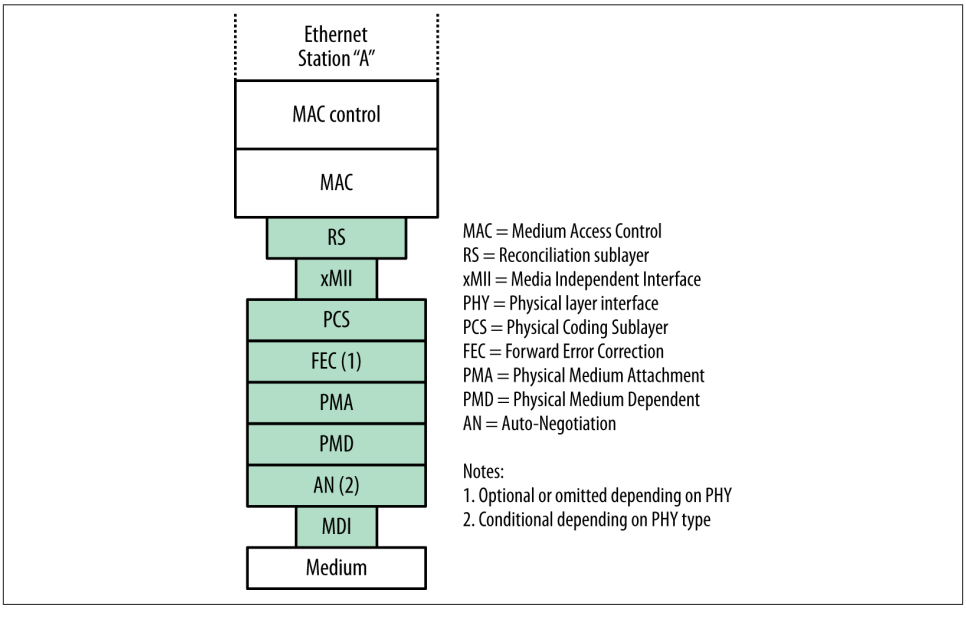


Figure 7-2. Ethernet physical layer elements

Also included in the set of specifications involved in putting Ethernet signals onto the medium is a reconciliation sublayer (RS), which is defined in the standard as a "mapping function that reconciles the signals at the Media Independent Interface (MII) to the

Media Access Control (MAC)–Physical Signaling Sublayer (PLS) service definitions."[1] In other words, the RS is a logical interface, provided to standardize the mapping for the signals carried between the MAC layer and the physical signaling layers.

These sublayers and other elements define the complex set of signaling standards that can coexist in a given Ethernet interface. For example, a modern 10 Gb/s Ethernet interface chip may support 100 Mb/s, 1 Gb/s, and 10 Gb/s operation over copper or fiber optic media systems. To do that, the chip is designed to provide multiple signaling systems over multiple internal signal paths, any of which can be configured internally in the chip to end up at the physical attachment to the medium.

This complexity is hidden from the user; all that you see is the eight-pin RJ45 copper port on a switch or computer, or a fiber optic transceiver that either is built in or that you insert into the Ethernet port on a switch or computer, depending on what the vendor chooses to provide in the way of options. The various signal encoding systems and logical interfaces are inside the port and are part of the Ethernet interface chipset. Which of the supported media systems and speeds is used at any given time may be chosen automatically by Auto-Negotiation, or configured manually by you in interaction with the interface management software on the switch or computer that you are connecting to the Ethernet.

Ethernet Signal Encoding

Signal encoding is a means of combining both clocking and data information into a self-synchronizing stream of signals sent over a media system. Each media system presents a certain challenge to the standards engineers in terms of sending Ethernet signals that can make it from one end of the cable to another.

As higher-speed Ethernet systems have evolved, more complex block encoding schemes have been developed. All of these signaling systems have the same set of goals. First, they must include sufficient clocking information along with the signals to ensure that the signal decoding circuitry can function correctly. Other goals include ensuring that the error rate is kept very low, and that the Ethernet frame data has a very high probability of surviving its trip over the media system.

Baseband Signaling Issues

The baseband signaling used in Ethernet media systems puts the bits of the Ethernet frame onto the link as a sequence of pulses or data symbols. These signals are reduced in amplitude and distorted by the electrical or photonic effects of being carried over the cable by the time they reach the other end of the link. The receiver's job is to correctly

1. IEEE Std 802.3-2012, paragraph 1.4.341, p. 38.

detect each signal pulse as it arrives and to decode the correct bit value before transferring the information to the receiving MAC.

Electrical and digital filters and pulse-shaping circuits are used to restore the size and shape of the received waveforms; other measures are also taken to ensure that the received signals are sampled at the correct time in the pulse period and at the same rate as the transmitting clock. Because the transmitting and receiving clocks are on separate devices, synchronizing the clocks is accomplished by sending clocking information with the data. The receiving device synchronizes to the incoming data stream using clocking information encoded in the data, which makes it possible to correctly decode the received data. Clock recovery requires sufficient signal level transitions in the incoming signals to make it possible for the receiving device to correctly identify symbol boundaries.

The earliest encoding scheme was Manchester encoding, used to transmit 10 Mb/s signals. Manchester encoding provides a signal transition in the middle of each bit symbol, which is used by the receiver as clocking information to enable it to stay synchronized with the incoming signals and to correctly decode the ones and zeros being sent.

The Manchester encoding system is inefficient, however, requiring two transitions to represent each bit in a string of all ones or zeros, and this presents difficulties for signaling over copper cables at higher speeds. This led to the adoption of more complex signal encoding schemes for higher-speed Ethernet systems over copper cabling that could transmit a given bit rate with fewer signal transitions than Manchester would require. On the other hand, most of the fiber optic media systems use simple encoding schemes, because fiber optic cables can sustain higher-frequency signaling than copper media and are not susceptible to electrical effects like baseline wander.

Baseline Wander and Signal Encoding

Baseline wander is an issue that arises when sending and receiving higher-speed electrical signals over copper cabling. The signal-receiving circuits may lose synchronization if the data being sent remains constant and provides no transitions to detect (e.g., during long strings of zeros). More complex encoding schemes make it possible to manage this effect.

Baseline wander occurs because the copper Ethernet media systems are electrically coupled with transformers to the receiving electronics to maintain electrical isolation. This provides a measure of safety in case the copper cabling should accidentally have a high voltage placed on it due to an electrical fault in the cabling system.

However, transformer coupling also allows the average signal level to vary, and if, for example, a long series of zeros containing no signal transitions is sent, the signal level can drop below the voltage threshold used to detect a one or a zero, causing an erroneous

value to be detected. To avoid this issue, faster Ethernet systems adopted a variety of techniques to improve signal transmission and recovery.

Advanced Signaling Techniques

To help avoid signal errors, higher-speed Ethernet systems employ a variety of techniques. These include:

Data scrambling
: The bits in each byte are scrambled in an orderly and recoverable fashion, to ensure that there are no long series of zeros or ones being transmitted, increasing "transition density." This avoids baseline wander and makes it easier to detect the clocking information in the stream of symbols being sent over the medium.

Expanded code space
: More signaling codes are added in this approach to represent both data and control symbols, such as start of stream and end of stream signals. This helps improve frame detection and error detection.

Forward error correcting codes
: This approach adds redundant information to the transmitted data, so that some types of transmission errors can be detected and corrected during frame transmission.

Continually increasing the speed of Ethernet has meant that cabling and connector technologies have also had to evolve to support the higher speeds. While the twisted-pair system has standardized on the use of the eight-position (RJ45) socket and plug connectors, the signal-handling quality of the twisted-pair cabling and connectors has steadily improved to support the higher signaling speeds. The fiber optic system has also evolved faster cabling and a variety of different connector types used to connect an Ethernet interface to the cabling.

Ethernet Interface

In the earliest days of Ethernet, the network interface card (NIC) was a fairly large circuit board covered with chips connected together to implement the necessary functions. Nowadays, an Ethernet interface is typically contained in a single chip, or even a portion of a larger "system-on-chip," that incorporates all of the required Ethernet functions, including the MAC protocol. Ethernet interface chips are designed to keep up with the full rate of the Ethernet media systems that they support.

However, the Ethernet interface is only one of an entire set of entities that must interact to make network services happen. Various elements have an effect on how many Ethernet frames a given Ethernet switch or desktop or server computer can send and receive within a specified period of time. These include the speed with which the switch or

computer system can respond to signals from the Ethernet interface chip, the amount of available port buffer memory for storing frames, and the efficiency of the interface driver software.

This is an important point to understand. All Ethernet interface chips are capable of sending and receiving a frame at the full frame rate for the media systems that they support. However, the total performance of the system is affected by the power of the computer's CPU, the speed of internal signaling pathways linking the CPU with the Ethernet interface, the amount of buffer memory, and the quality of the software that interacts with the Ethernet interface. None of these elements are specified in the Ethernet standard.

If the computer system is not fast enough, then Ethernet frames may not be acknowledged or received. When that happens, the frames are dropped or ignored by the Ethernet interface. This is acceptable behavior as far as the standard is concerned, because no attempt is made to standardize computer performance.

These days, most computers are capable of sending and receiving a constant stream of Ethernet frames at the maximum frame rate of a 10 Mb/s, 100 Mb/s, or 1 Gb/s Ethernet system. However, slower computers, such as embedded servers with low-power CPUs and slow internal communication paths, may not be able to keep up with the full frame rate on the Ethernet systems to which they are attached.

Higher-Speed Ethernet Interfaces

It's possible for a computer system to use a significant percentage of its CPU power to receive or transmit the full frame rate on high-speed Ethernet systems. You should be aware of these performance issues, and not assume that a given system can be connected to high-performance Ethernet links without issues.

If a machine that is working hard to keep up with a 1 Gb/s Ethernet channel is connected to a 10 Gb/s Ethernet, it will not suddenly be able to go 10 times faster—10 Gb/s Ethernet speeds and frame rates push the limits of even high-performance computer systems. As of this writing, a number of the desktop and server computers currently on the market cannot keep up with the full frame rate of a 10 Gb/s Ethernet channel.

While the signal paths inside high-performance servers may have enough bandwidth to keep up with a 10 Gb/s Ethernet channel, the network interface may still require the use of special software to help speed up processing of network protocol software, and to improve data rates between the server's CPU and the 10 Gb/s Ethernet interface. Some vendors supply interfaces that provide high-level protocol packet processing onboard, to speed the flow of packets between the computer and the network. Other approaches include more sophisticated interface drivers capable of buffering several packets before interrupting the computer's CPU. Yet another approach is the use of direct memory access techniques to manage packet flow in and out of the interface.

Next we'll look at how the media signaling can be modified to reduce energy requirements.

Energy Efficient Ethernet

Now that we've seen how the media signaling components are organized and work together to send Ethernet signals from one station to another, this is a good place to describe how Energy Efficient Ethernet (EEE, pronounced "triple E") can modify those signals to save power. EEE is an optional standard that currently applies to twisted-pair Ethernet media systems and also to the Ethernet standards used for sending signals over backplanes in devices such as chassis switches. Future extensions to the standard are expected to include more media systems.

One way to describe EEE is to compare it with the operational modes in standard and hybrid cars. When a hybrid car stops at an intersection, the gasoline engine is shut off to save power, and it is restarted when you press on the accelerator. Prior to the EEE standard, all Ethernet ports operated like standard cars with an engine that continued to run even when the car was stopped. The EEE standard makes it possible for ports to operate more like hybrid cars, and automatically shuts off some interface functions, minimizing the power needed to operate an Ethernet port until there is data to send.

Researchers at the University of South Florida initially raised the issue of saving power on Ethernet links, with a proposal called Adaptive Link Rate.[2] The researchers noted that hundreds of millions of Ethernet links were operating at full signaling speeds 100% of the time and consuming electrical power to do so, even though the signaling was only being used to send IDLE symbols much of the time.

The researchers found that many Ethernet links were operating at relatively low utilization rates, and that there were significant periods of time when no user data was being sent. For example, many desktop computers are not used after the working day ends, yet they continue to send Ethernet signals at full speed over the link all night long, just to indicate that the link is idle.

The researchers cited a 2002 study that found commercial office and telecommunications equipment to account for about 2.7% of U.S. electricity consumption in 2000. That study found that the networking equipment alone in nonresidential U.S. office spaces —not including PCs, monitors, servers, and the like—consumed 6.4 trillion watt hours in 2000.

The researchers noted that varying the signaling rate on an Ethernet link from 1 Gb/s to 100 Mb/s made a difference of about 4 watts of power consumption per port. They

2. C. Gunaratne and K. Christensen, "Ethernet Adaptive Link Rate: System Design and Performance Evaluation," *Proceedings 2006 31st IEEE Conference on Local Computer Networks* (Nov. 2006): 28–35.

calculated that if the estimated 160 million Ethernet-connected PCs in the United States could operate their links at lower power levels when idle, the decreased power consumption could save $240 million a year.

Estimates published by Broadcom indicate that powering down networking ports when there is no data to send could reduce the energy required by the physical layer operations by up to 70% or more, making it possible to achieve overall savings of over 33% in the power required to operate an Ethernet switch.

IEEE EEE Standard

In response to these concerns, Energy Efficient Ethernet was developed over a period of several years and specified in the 802.3az supplement to the standard. The 802.3az supplement was approved as a standard on September 30, 2010, and was adopted as Clause 78 of the 2012 edition of the 802.3 standard. EEE provides a low power idle (LPI) mode of operation for media systems that use block encoded symbols. The standard also provides a lower-power version of the simpler 10 Mb/s Manchester encoded signaling.

The EEE system uses Auto-Negotiation to advertise EEE capabilities between link partners to determine whether EEE is supported, and to select the best set of parameters supported by both devices. When LPI mode is enabled, the Ethernet devices at both ends of a link can save power by effectively shutting off their transmitter and receiver circuits during periods of low link utilization, or by using the lower-power version of Manchester signaling if operating at 10 Mb/s.

EEE signals are used to transition to a lower level of power consumption; this is accomplished without changing the link status and without dropping or corrupting frames. The transition time into and out of low power consumption mode is designed to be small enough to be ignored by upper-layer protocols and applications. The goal was to avoid causing any noticeable delays when the EEE system is operating.

EEE media systems

EEE is currently supported over 100BASE-T, 1000BASE-T, and 10GBASE-T twisted-pair media systems. For operation over electrical backplanes, EEE also supports the 1000BASE-KX, 10GBASE-KX4, and 10GBASE-KR backplane media standards. The EEE standard also defines a reduced-power version of the 10 Mb/s signaling, called 10BASE-Te. The 10BASE-Te system fully interoperates with existing 10BASE-T transceivers over 100 meters of class D (Category 5) cabling, to provide a reduction in power consumption for systems operating at 10Mb/s. The 10BASE-T system does not use block encoding so it cannot support the EEE protocol.

Efforts to extend EEE capability into other media standards are underway, with work proceeding in the 40 and 100 Gb/s standards for backplanes and copper cabling. Efforts are also being made to extend EEE capability to fiber optic media systems.[3]

EEE Operation

EEE saves power by switching off functions in the Ethernet interface when no data needs to be transmitted or received. The decision as to whether the link should enter or exit low power idle mode is made by the interface software, based on whether or not there is Ethernet frame data that needs to be sent.

EEE operates over a link between two stations, and only for stations that are in the full-duplex mode of operation. Both stations must support EEE, or LPI mode cannot be enabled. When the stations first come up over the link, they use Auto-Negotiation to advertise their EEE capabilities.

Once the link partners have determined that they both support EEE, then the stations can use LPI signaling to indicate that there is a break in the data being sent, and that the link can go quiet until there is data to send. The EEE protocol uses a signal that is a modification of the normal IDLE symbol that is transmitted between frames on systems with complex encoding.

Figure 7-3 shows the LPI signals that are sent and received by the PHY. When the controller software determines that there is no data to send and that it can enter LPI mode, an LPI TX Request is sent. Following reception of the request, the PHY transmits LPI symbols for a defined period (T_s = time to sleep), the transmitter stops signaling, and the link enters LPI mode. This is the behavior for most of the supported media types.

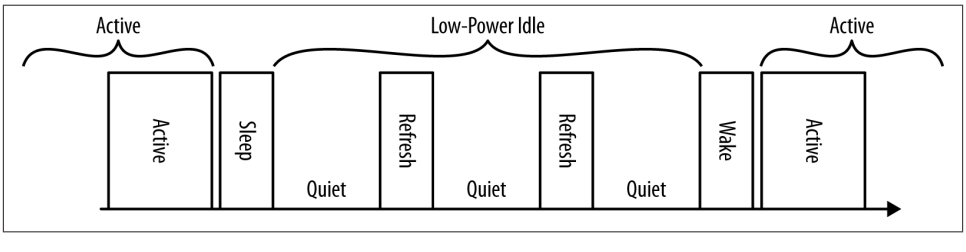


Figure 7-3. LPI signals

However, in the 1000BASE-T system, which uses a master-slave method for synchronizing signals sent over the link, the PHY operation for initiating LPI mode is symmetric.

3. Wael William Diab, "The Power and Promise of Energy Efficient Ethernet (EEE): A State of the Union Address," Ethernet Alliance Blog, January 11, 2013.

Only after the local 1000BASE-T PHY both transmits sleep symbols to *and* receives sleep symbols from the link partner does the PHY enter the quiet mode.

EEE states

While in quiet mode, the local PHY periodically transmits refresh signals. The refresh pulse that is periodically sent while the link is in LPI mode serves the same purpose as the link pulse in the 10BASE-T media system, and maintains link state. The frequency of the refresh symbols, which are sent multiple times per second, also prevents any situation where one link partner is disconnected and another connected without causing a link fail event. This maintains compatibility with any mechanisms that depend on continuous connectivity and that need to know when a link is disconnected.

The refresh signals are also intended to provide enough signaling over the cable to update adaptive filters and timing circuits in order to maintain link signaling integrity. The specific timing between refresh signals and the type of signal used for refresh varies according to each media type, to ensure that the specific media signaling requirements are being met to maintain a stable link signal while idle (which also ensures that the link can rapidly return to full operation).

This "quiet/refresh" cycle continues until the controller software detects that there is data to send, at which point it sends a message to clear the LPI mode. In response, the transmitter begins sending normal IDLE symbols, and after a predetermined time called T_w (time to wake) the PHY enters the active state and resumes normal operation.

The EEE protocol allows a link to be reawakened at any time, and there is no minimum or maximum sleep interval. The default wake time for each type of media system (PHY) is designed to be similar to the time taken to transmit a maximum-length frame. For example, the worst-case wake time for 1000BASE-T is 16.5 µs (16.5 millionths of a second), which is roughly the same time it takes to transmit a 2000-byte Ethernet frame over that media system. The wake time is defined in the standard as T_{w_phy}, and the maximum time to reawaken a link is defined as $T_{w\_sys\_tx}$, or "the longest period of time the system has to wait between a request to transmit and its readiness to transmit."

Table 7-2 lists the wake times and maximum times to reawaken for some common media types. For two of the three media systems shown, the wake time and max time to reawaken are the same. The wake time on the 100BASE-TX system is 20.5 µs, and the maximum time to reawaken a 100BASE-TX link is 30 µs.

Table 7-2. EEE wake and reawaken timing

Media type	T_{w_phy}	$T_{w\_sys\_tx}$
100BASE-TX	20.5 µs	30 µs
1000BASE-T	16.5 µs	16.5 µs
10GBASE-T	7.36 µs	7.36 µs

Managing EEE

While the EEE standard defines how low power idle mode is communicated between stations on a link and how the PHYs can transition into and out of this mode, it does not define when LPI mode can be used. The determination of when to enter LPI mode is left up to the system. Each system is expected to determine the policy for when to enter LPI mode. The policies used may include:

Simplest policy
: When the transmit buffer is empty, wait a short time and then request LPI mode. When a frame arrives to send, reawaken the link.

Buffer and burst policy
: When the transmit buffer is empty, request LPI mode. When one or more frames arrive for transmission, wait until a large enough set of frames arrives or until a timer expires, and then reawaken the link.

Application-aware policy
: Monitor the transport or higher-layer communication software to understand when the link can be deactivated, or whether more packets should be expected soon.

Early EEE systems may support only the simplest policy. As vendors gain more experience with the system, it is expected that more complex policies will be developed. One area of interest is the power required for data center operations, and the development of energy management systems for centers with thousands of servers and network ports that could use EEE to realize energy savings across the set of data center ports.

EEE negotiation

The EEE protocol also provides a method for link partners to exchange LLDP packets, defined in IEEE 802.1AB, to negotiate wake times that are different than the defaults in the standard. The LLDP standard is already widely supported in networking equipment, making it possible to add wake timing negotiation without requiring that the switch support a new protocol. The LLDP-based negotiation ability allows a vendor to program equipment like a PC to go into a deeper sleep with more components powered down, resulting in greater power savings.

However, a deeper sleep also requires a longer time to wake, hence the need to renegotiate the wake timing. The wake timing can be negotiated separately for each direction over the link depending on the equipment at each end, so that wake times can be asymmetric.

Impact of EEE Operation on Latency

Latency is the time required for an Ethernet frame to transit the set of equipment between a sending computer and the receiving system. Latency includes the inevitable serialization delays caused by transmitting Ethernet frames one bit at a time over Ether-

net links. It also includes any time consumed by moving the Ethernet frame into and out of switch port buffers and across switching backplanes and switch fabrics.[4] The general goal is to minimize latency, so as to avoid any impacts on delay-sensitive applications, such as voice or video, where excessive packet delays may reduce voice quality or cause issues for video images.

The EEE protocol was designed to minimize the delays that occur when entering and leaving idle mode. The default wake times are similar to the amount of time required for transmitting a maximum-sized frame on the media system in question. This design keeps the impact on applications to a minimum, as a similar delay is incurred during the normal store and forward packet switching functions in Ethernet switches, in which the entire frame is read into port buffer memory (store) before being sent out the port, one bit at a time (forward).

However, some applications can be extremely sensitive to any extra latency. Some high-performance computing systems can be sensitive to the delay times experienced by interprocessor communications or synchronization traffic carried over Ethernet channels. Some financial trading applications also make strenuous attempts to minimize delay, using techniques such as cut-through switching to avoid spending any time on normal store and forward switching operations. These applications could be affected by the sleep and wake times incurred by EEE operation. In these cases, you may wish to disable EEE operation over the links involved.

Normal network traffic, including IP video, telephony, and telepresence, are designed to work over normal networks and typically have a built-in latency tolerance of anywhere from 1 to 10 milliseconds. This is quite a lot larger than the microseconds required for default EEE sleep and wake operations. Therefore, EEE operations using default times should have no impact on these applications.

EEE Power Savings

The EEE system makes it possible to achieve significant power savings, using an automatic system that causes links to dynamically enter and leave a low power idle mode depending on link traffic. The system is designed to operate invisibly, and does not require any user intervention to function. Now that Ethernet interface vendors are shipping interface chips that include EEE support, a link with EEE capability on both ends can automatically negotiate the use of EEE, saving power when there is no data to send.

4. A definition of data communications latency is provided in RFC 1242, and a method for measuring switch latency is provided in RFC 2544. The QLogic "Introduction to Ethernet Latency" white paper describes latency testing in detail.

EEE power savings in an interface

Intel has published a measurement of the power savings that can be realized using its 82579 Gigabit Ethernet interface chip, which supports 100 and 1000 Mb/s operation.[5] It shows the power consumed when the link is sending a frame, when sending normal IDLE symbols at full speed in the normal operating mode, and when in LPI mode.

Table 7-3 shows that EEE can reduce the power required to maintain idle operations on 1000BASE-T links by 91% when there are no frames to send, and by 74% on 100BASE-T links. While the actual amounts of power being saved are in the milliwatts, this adds up in a major way across thousands of ports at a given site and hundreds of millions of ports across the worldwide networks.

Table 7-3. EEE power savings for the 82579 interface chip

Media speed	Link state	Power consumed in milliwatts (mW)
1000 Mb/s	Active	619
1000 Mb/s	Idle	590
1000 Mb/s	LPI	49
100 Mb/s	Active	315
100 Mb/s	Idle	207
100 Mb/s	LPI	53

EEE power savings in a switch

Cisco Systems tested power consumption in one of its Catalyst 4500 switches with EEE enabled. The test consisted of connecting 384 ports to a stream of traffic being generated to simulate a bursty traffic pattern, of the kind typically seen being generated by desktop computers.[6]

The generated packet bursts were separated by 100 milliseconds, with 100,000 64-byte packets in each burst. Each port was connected to the adjacent port with a cable, such that traffic injected into port 1 was transmitted onto port 2, which was connected via cable from port 2 to port 3. Port 3 transmitted onto port 4, which was connected via cable to port 5, and so on. The goal was to see what kind of power savings EEE could achieve on 384 ports that were all carrying traffic that roughly simulates normal desktop user activity. Not counting the ports used to inject the packets and send the packet stream to the test equipment, there were a total of 191 EEE-enabled links connecting the ports together.

5. Jorden Rodgers (JordanR), "Energy Efficient Ethernet: Technology, Application and Why You Should Care," Wired Ethernet, May 5, 2011.

6. Cisco Systems, Inc. and Intel, "IEEE 802.3az Energy Efficient Ethernet: Build Greener Networks," 2011.

The power consumed by the switch was measured before and after enabling EEE. Prior to enabling EEE, the switch consumed 892 watts while running the packet test through all ports. After enabling EEE, the power consumption dropped to 751 watts total. The power reduction of 141 watts resulted in an average power saving per link of 0.74 watts when EEE was active. This test achieved roughly a 15% reduction in power consumed on 191 links, showing that EEE can achieve significant power savings even with continuous bursts of activity on all ports.

CHAPTER 8

10 Mb/s Ethernet

This chapter describes the signaling and media components used in the 10 megabit per second Ethernet media systems. We also provide the basic configuration guidelines for both copper and fiber optic cable segments operating at 10 Mb/s.

The original 10 Mb/s Ethernet system was based on coaxial cable segments. There were two kinds of coaxial cable used: the original "thick coax" system, with a cable that was approximately one half inch in diameter, and the "thin coax" system, which used a cable that was roughly one quarter inch in diameter. The thick and thin coax systems used an external *medium attachment unit* (MAU), also known as a *transceiver*, to connect the Ethernet interface to the cable.

The connection between the interface and the MAU was called an *attachment unit interface* (AUI), also known as a *transceiver cable*. The thin coax interfaces often included an internal transceiver as well, and provided a BNC coaxial connector on the interface for direct connection to the coax, saving the cost and complexity of the external transceiver connection. The external transceiver (MAU) and transceiver cable (AUI) that were part of the coaxial cable systems are no longer used.

The coaxial cable systems are obsolete; all modern Ethernet systems are based on twisted-pair and fiber optic cables. The 10 Mb/s system allows for both, although since the adoption of higher-speed systems the 10 Mb/s fiber optic system is no longer used, making the 10BASE-T system based on twisted-pair cabling the most widely used version of 10 Mb/s.

10BASE-T Media System

The 10BASE-T system was the first popular twisted-pair Ethernet system. Its invention in the late 1980s led to the widespread adoption of Ethernet for desktop computers. The 10BASE-T system was originally designed to support the transmission of 10 Mb/s Ethernet signals over "voice-grade" Category 3 twisted-pair cables. However, the vast

majority of twisted-pair cabling systems in use today are based on Category 5/5e twisted-pair cables, or better. These newer cables have higher-quality signal-carrying characteristics that work very well with the 10BASE-T system.

10BASE-T Ethernet Interface

A typical twisted-pair Ethernet interface includes an eight-pin connector (RJ45) with a built-in transceiver that is used to make a direct connection to the twisted-pair segment. The transceiver operates at 10, 100, or 1000 Mb/s.

Signals for 10BASE-T operation travel over two pairs of a twisted-pair link, making it possible for 10BASE-T to operate on telephone-grade cabling that provides only two signal pairs.

Signal Polarity and Polarity Reversal

The transmit or receive data signals used for 10BASE-T signals over a twisted-pair segment are polarized, with one wire of each wire pair carrying the positive (+) signal, and the other carrying the negative (–) signal. Many 10BASE-T transceivers support an optional feature called *polarity reversal*, which can automatically detect and correct wiring errors that result in incorrect polarity in a given wire pair.

Polarity reversal refers to swapping the position of the two wires within a given wire pair. This is different from a wiring crossover error, which may, for example, involve swapping the position of wire pair 2 with wire pair 3.

10BASE-T Signal Encoding

Signals sent over 10 Mb/s media systemsuse a relatively simple encoding scheme called Manchester encoding, which is named for its origin at Manchester University in England. Manchester encoding combines data and clock signals into *bit symbols*, which provide a clocking transition in the middle of each bit. As shown in Figure 8-1, each Manchester-encoded bit is sent over the network in a *bit period* that is split into two halves, with the polarity of the second half always being the reverse of the first half.

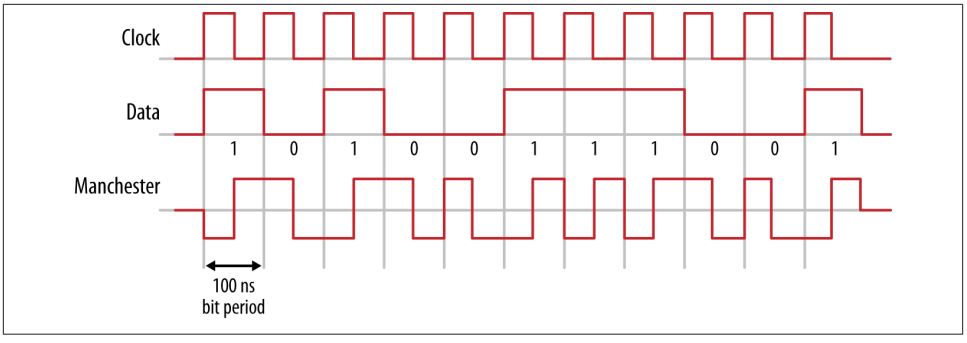


Figure 8-1. Manchester signals over 10BASE-T

The rules for Manchester encoding define a 0 as a signal that is high for the first half of the bit period and low for the second half. A 1 is defined as a signal that is low for the first half of the bit period and high for the second half. Figure 8-1 shows a station sending the bit pattern 10100111001. The encoded signal is the exclusive OR (XOR) of the clock signal and the data.[1]

Manchester encoding provides a clock transition in each digital bit sent. This transition is used by the receiving station to synchronize itself with the incoming data. While Manchester encoding makes it easy for a receiver to synchronize with the incoming signal and to extract data from it, a drawback of the scheme is that it is inefficient in its use of bandwidth, as it requires two transitions to send a single bit when a series of ones or zeros is being transmitted. Stated differently: the baud rate is twice the bit rate.

Physical line signaling

10BASE-T transceivers are designed to send and receive signals over four wires of a twisted-pair cable segment: one pair of wires for transmitting data and another pair for receiving data. The 10BASE-T line signals are sent over the twisted-pair wires as balanced differential voltages. In each wire pair, one wire is used to carry the positive amplitude of the differential signal (from 0 volts to +2.5 volts), and one wire carries the negative amplitude of the signal (from 0 volts to –2.5 volts).

Differential signaling provides its own zero reference point, around which the electrical signals swing positive or negative. This avoids any requirement to reference the signals on a 10BASE-T segment to a common ground level shared by the equipment at both ends. The 10BASE-T system does not reference the signals to a common ground; thus the system is isolated from variations in ground voltage that can occur in a twisted-pair cabling system. That, in turn, eliminates problems with ground currents and improves the reliability of the system.

1. An exhaustive description of the XOR logical function can be found on Wikipedia.

10BASE-T Media Components

The following set of media components are used to build a 10BASE-T twisted-pair segment:

- Unshielded twisted-pair (UTP) cable, Category 3 or better
- Eight-position RJ45-style modular connector

UTP cable

The 10BASE-T system operates over two pairs of UTP wires; one pair receives data signals into the station or hub port, and the other pair is used for transmitting data signals from the station or hub port. The 10BASE-T standard was designed to accommodate twisted-pair cabling systems based on ordinary voice-grade telephone wire rated to meet the TIA/EIA Category 3 specifications (see Chapter 15). The target length in the standard for a 10BASE-T segment based on voice-grade cabling and components is 100 meters (328 feet). More details on installing and using twisted-pair cables and connectors can be found in Chapter 16.

A 10BASE-T segment can be longer than 100 meters, as long as the signal quality specifications are met. Most of the time this will not matter, as the vast majority of all office and work areas are within 100 meters of a telecommunications closet. However, on occasion, you may need a 10BASE-T segment that is longer than 100 meters to reach equipment that's further away from the closet. Note that a segment longer than 100 meters is unlikely to support higher-speed media systems, and will require you to manually configure the speed to ensure that it does not exceed 10 Mb/s. The next section will discuss ways to increase the length of your 10BASE-T segments.

10BASE-T segments longer than 100 meters. The major limiting factor on a 10BASE-T segment is the strength of the signal, or *signal attenuation*. The receiver circuit in a typical 10BASE-T transceiver has a signal squelch level set at 300 millivolts (mV), which helps prevent the signals induced by electrical noise or signal crosstalk between wire pairs from becoming a problem by limiting the level at which signals are received. Once a signal sinks below this level, it will not be received by a 10BASE-T transceiver. With this approach, signals below 300 mV are simply ignored. However, this also means that when signal attenuation over a long segment lowers the real signal level to below 300 mV, the segment will stop working.

The maximum signal attenuation allowed in the specifications for a 10BASE-T segment is 11.5 decibels (dB), as measured from one end of the segment to the other with a cable testing device. A typical Category 5 cable has an attenuation of 10 dB per 500 feet at 10 MHz frequencies. Therefore, 500 feet of this kind of twisted-pair cable would use up the majority of the 11.5 dB signal loss that is allowed on a 10BASE-T segment.

You can expect that at least 1.5 dB of the loss budget will be used up by the signal losses that occur in RJ45-style connectors, patch panels, and patch cables. Taken all together, even if you use Category 5 cable, it will be difficult to achieve a segment any longer than approximately 150 meters (roughly 490 feet) while staying within the 10BASE-T signal quality specifications found in the standard.

A 10BASE-T segment will be more likely to function over distances greater than 100 meters if you use high-quality, low-attenuation twisted-pair cable and keep the number of connectors and patch panels to a minimum. The further you go beyond 100 meters, however, the greater the total amount of signal attenuation will be.

Twisted-pair impedance rating. For best results, you should use twisted-pair cable with a 100 ohm characteristic impedance rating. However, the standard notes that it is possible to build 10BASE-T segments using twisted-pair cable with a 120 ohm characteristic impedance. If you must use cable with a 120 ohm impedance, then you should check with the vendor of your equipment to see if the equipment is designed to function adequately with twisted-pair cables at that impedance level.

Eight-position RJ45-style jack connectors

The 10BASE-T media system uses two pairs of wires, which are terminated in an eight-position (RJ45-style) connector. This means that four pins of the eight-pin connector are used. Table 8-1 lists the 10BASE-T signals used on the eight-pin connector.

Table 8-1. 10BASE-T eight-pin connector signals

Pin number	Signal
1	TD+ (transmit data)
2	TD – (transmit data)
3	RD+ (receive data)
4	Unused
5	Unused
6	RD – (receive data)
7	Unused
8	Unused

A typical twisted-pair segment will have all eight wires connected to the RJ45-style connector in the standard configuration used for structured cabling systems, even though the 10BASE-T media system only uses four of the eight wires.

The TIA/EIA structured cabling standard recommends installing two twisted-pair cables for each office: one for data service and another for telephone or other service. A conservative design reserves a four-pair cable for data service, uses a cable rated to meet

the Category 5e or better specifications, and connects all eight wires of the cable. That way, the network can provide 10, 100, and 1000 Mb/s service.

Connecting a Station to 10BASE-T Ethernet

Now that we've seen the components that make up a 10BASE-T Ethernet system, let's look at how these components are used to connect a station to a twisted-pair segment.

Figure 8-2 shows a computer with a built-in Ethernet interface that supports 10BASE-T operation. The interface has an RJ45 connector, to which the twisted-pair segment is directly attached. A signal crossover is needed in the signal path of each twisted-pair segment to ensure that the Ethernet signals are connected properly. Signal crossover for twisted-pair cables and connectors is described in Chapter 16.

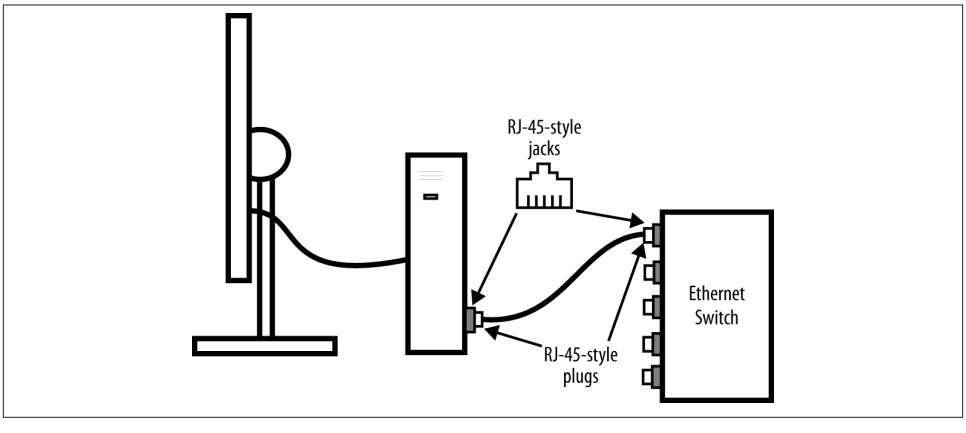


Figure 8-2. Connecting a station to a 10BASE-T link

10BASE-T Link Integrity Test

A transceiver operating in 10BASE-T mode continually monitors the receive data path for activity as a means of checking whether the link is working correctly. The transceiver also sends a link test signal to verify the integrity of both twisted-pair links. The link signal is only sent when there is no other data on the network, so there is no performance impact caused by sending link signals. Vendors can optionally provide a link light on the Ethernet interface. If the link lights on the Ethernet ports at both ends of a segment are lit when you connect a segment, then you have an indication that the segment is wired correctly.

The presence of a link light at both ends indicates that the basic transceiver functions are working, and that a signal path exists between the transceivers. It's important that the link lights on both interfaces be lit, because the lights indicate whether both signal paths between the two devices are wired correctly.

The link test signal pulse operates more slowly than actual Ethernet signals, so the presence of link lights won't guarantee that Ethernet signals will work over the segment. Odds are good that a correctly wired segment will work, but if the signal crosstalk on the segment is too high, then it may not work despite the presence of the link lights.

10BASE-T Configuration Guidelines

The Ethernet standard contains guidelines for building a twisted-pair segment that supports 10BASE-T operation. Table 8-2 lists the guidelines for a 10BASE-T segment.

Table 8-2. 10BASE-T single segment guidelines

Media type	Maximum segment length	Maximum number of connections (per segment)
Twisted-pair 10BASE-T	100 m (328 feet)	2 (one at each end)

As noted, while 100 meters is the target length set in the standard for 10BASE-T segments based on Category 3 (voice-grade) cable and components, 10BASE-T segments may be longer while still meeting the electrical specifications in the standard, depending on the quality of the twisted-pair segment.

There is no minimum length specification for a 10BASE-T segment. In practice, you can purchase ready-made patch cables as short as one foot, which can be used to connect 10BASE-T equipment together. However, you may find that if you want to test the cable with a handheld cable tester, there may be a minimum cable length (typically two meters) that the tester requires for an accurate test of cable parameters.

Fiber Optic Media Systems (10BASE-F)

In this section, we discuss the evolution of the 10BASE-F system. We also discuss the signaling and media components used in 10BASE-F.

The 10 Mb/s fiber optic media segments are no longer widely used, having been superseded by faster media systems. Nonetheless, 10BASE-FL links were sold and used for many years, and it is still possible to purchase 10BASE-FL transceivers. We will describe the 10BASE-F standard in this chapter for the sake of completeness.

The 10BASE-F fiber optic media system uses pulses of light to send Ethernet signals, which has several advantages. For one thing, a fiber optic link segment can carry Ethernet signals for considerably longer distances than metallic media can. Fiber optic media, widely used as the backbone cabling in structured cabling systems, allows you to link Ethernet switches located on each floor of a building with a media system that can travel longer distances than twisted-pair segments. Fiber optic media can also support higher-speed Ethernet systems. This means that the fiber optic media you install to support 10 Mb/s Ethernet operation can also be used for faster Ethernet systems.

Old and New Fiber Link Segments

Two 10 Mb/s fiber optic link segment types were standardized: the original Fiber Optic Inter-Repeater Link (FOIRL) segment, and the newer 10BASE-FL segment. The original FOIRL specification described a link segment of up to 1,000 meters for use between half-duplex signal repeaters only. Later, a new standard called 10BASE-F was developed, which provided a set of fiber media specifications including a link segment to allow direct attachments between switch ports and stations. The 10BASE-F standard includes three fiber optic segment types:

10BASE-FL
: The *fiber link* (FL) standard replaced the older FOIRL link segment, and 10BASE-FL signaling equipment was designed to interoperate with existing FOIRL-based equipment. 10BASE-FL provided a fiber optic link segment that could be up to 2,000 meters long, provided that the segment only used 10BASE-FL devices. If older FOIRL equipment was mixed with 10BASE-FL equipment, then the maximum segment length could only be 1,000 meters.

  The 10BASE-FL specs were the most widely used portion of the set of 10BASE-F fiber optic specifications, and 10BASE-FL equipment was available from a large number of vendors.

10BASE-FB
: The 10BASE-FB specification described a synchronous signaling *fiber backbone* (FB) segment. This media system allowed multiple half-duplex Ethernet signal repeaters to be linked in series, exceeding the limit on the total number of repeaters that could be used in a given 10 Mb/s Ethernet system. 10BASE-FB links were attached to synchronous signaling repeater hubs and used to link the hubs together in a half-duplex repeated backbone system that could span longer distances. Individual 10BASE-FB links could be up to 2,000 meters in length. However, the 10BASE-FB system was not widely adopted. For the first few years after the standard was developed, equipment was available from a few vendors, but this equipment is no longer sold.

10BASE-FP
: The *fiber passive* (FP) standard provided the specifications for a "passive fiber optic mixing segment." This was based on a non-powered device that acted as a fiber optic signal coupler, linking multiple computers on a fiber optic media system. According to the standard, 10BASE-FP segments could be up to 500 meters long, and a single 10BASE-FP fiber optic passive signal coupler could link up to 33 computers. This system does not appear to have been developed by any vendor, and equipment based on this standard doesn't exist.

Next, we will describe the 10BASE-FL fiber link segment and the older FOIRL segment, because these segments were the most widely used 10 Mb/s fiber optic segments.

10BASE-FL Signaling Components

The following signaling components may be used in the 10BASE-FL system to send and receive signals over the media system:

- An Ethernet interface equipped with a 10BASE-FL transceiver. A 10BASE-FL connection was most often provided as an external transceiver attached to a 15-pin AUI connector on the interface.
- A transceiver cable, also called an *attachment unit interface* (AUI). Note that transceiver connections to external transceivers are no longer used.
- An external 10BASE-FL transceiver, also called a *fiber optic medium attachment unit* (FO-MAU). In modern equipment, 10BASE-FL transceivers are built into the ports on switches and media converters.

10BASE-FL Ethernet Interface

Fiber optic connections are often used as uplink connections for switches, as described in Chapter 19. Uplinks on modern switches usually provide connections to high speed fiber optic media systems, such as 1 and 10 Gigabit Ethernet.

However, you can still purchase some 10BASE-FL components, including 10BASE-FL "media converters" which consist of 10BASE-FL fiber optic connectors and an RJ45 connector on the same device. The two media types are connected together in the media converter electronics, which are often nothing more than a simple two-port Ethernet switch on a chip. This provides a way to convert between a 10BASE-FL segment and a 10BASE-T segment, making it possible to extend the distance between two 10BASE-T devices with a 10BASE-FL link.

10BASE-FL Signal Encoding

Signals sent over the 10BASE-FL media system use the Manchester encoding system described earlier.

Physical line signaling

10BASE-FL transceivers send and receive signals as light pulses over a fiber optic segment that consists of two fiber optic cables: one cable for transmitting data and one cable for receiving data. This is done using a very simple line signaling scheme called *Non-Return-to-Zero* (NRZ). NRZ results in a light pulse being transmitted for a logical one (1) and no light pulse for a logical zero (0).

Signals are sent over a 10BASE-FL segment by turning the light on and off to indicate the Manchester-encoded signals representing ones and zeros. The Manchester encoding

ensures that there are enough logic transitions in the signal stream to provide clocking information for the signal decoding circuits.

10BASE-FL Media Components

The following media components are used to build a 10BASE-FL fiber optic segment:

- Multimode fiber optic cable
- Fiber optic connectors

We'll look at the characteristics of these cables and connectors in the following section, and provide the basic configuration guidelines for a single 10BASE-F segment.

10BASE-FL Fiber Optic Characteristics

The fiber optic cable specified in the standard for a fiber link segment consists of a graded-index multimode fiber cable (MMF) with a 62.5 micron (μm) fiber optic core and 125 μm outer cladding. The shorthand designation for this kind of fiber is 62.5/125. This is also known as "OM1" cable, based on the ISO/IEC 11801 standard for telecommunication cables.

Each fiber optic link segment requires two strands of fiber, one to transmit data and one to receive data. There are many kinds of fiber optic cables available, ranging from simple two-strand jumper cables with a plain PVC outer jacket material on up to large interbuilding cables carrying many fibers in a bundle. More details on fiber optic cables and connectors can be found in Chapter 17.

The 10BASE-FL fiber link system uses LED transmitters operating at a wavelength of 850 nanometers (nm).[2] The optical loss budget for a 10BASE-FL link segment must be no greater than 12.5 dB. As a very rough rule of thumb, a length of OM1 62.5/125 fiber optic cable carrying 10 Mb/s signals and operating at a wavelength of 850 nm will have roughly 3–4 dB loss per 1,000 m.

The loss could be higher depending on the number and quality of the splices in the cable. You can also expect anywhere from 0.5 dB to around 2.0 dB of loss per fiber optic connection point, depending on how well the connection has been made.

The older FOIRL segment standard specified the same type of 62.5/125 fiber optic cable, and had the same 12.5 dB optical loss budget. The 10BASE-FL specifications were designed to allow backward compatibility with existing FOIRL segments. The major difference is that the 10BASE-FL segment may be up to 2,000 m in length if 10BASE-FL

2. A nanometer is one billionth of a meter.

equipment is used on both ends of the segment, while the FOIRL segment was limited to a maximum of 1,000 m.

Alternate 10BASE-FL Fiber Optic Cables

Over the years, a variety of fiber optic cables have been used in various proprietary networks and cabling systems. The IEEE 802.3 standard states that these cables may also be used as alternates to the standard 62.5/125 cable in a 10BASE-FL link. Cables with a 50 m fiber optic core and 125 μm outer cladding (50/125), as well as 85/125 and 100/140 cables, were considered alternate cables when the standard was first published. The standard notes that details for the use of alternative cables are not provided, and their use may reduce the maximum achievable distance of a segment.

The difficulty here is that the use of these cables causes a mismatch between the size of the fiber optic core on the alternate cables and the standard 62.5 μm size of the receivers and transmitters on 10BASE-FL equipment. While the alternate cables can be terminated in ST fiber optic connectors and connected to 10BASE-FL equipment, there can be a significant loss in signal due to the mismatch in size. When using cables with a core size other than 62.5 μm, the losses due to mismatch can be as high as 5 or 6 dB, or even more. In that case, the total length of the segment must be reduced to accommodate the higher losses at the connection points at each end of the segment.

Fiber Optic Connectors

The fiber connector used on 10BASE-FL link segments is generally known as an ST connector, where ST stands for *straight tip*. The formal name of this connector in the ISO/IEC international standards is *BFOC/2.5*.

Figure 8-3 shows a pair of fiber optic cables equipped with ST plug connectors. The ST connector is a spring-loaded bayonet connector whose outer ring locks onto the connection. The ST connector has a key on an inner sleeve along with the outer bayonet ring. To make a connection, you line up the key on the inner sleeve of the ST plug with a corresponding slot on the ST receptacle. You then push in the connector and lock it in place by twisting the outer bayonet ring. This provides a tight connection with precise alignment between the two pieces of fiber optic cable being joined.

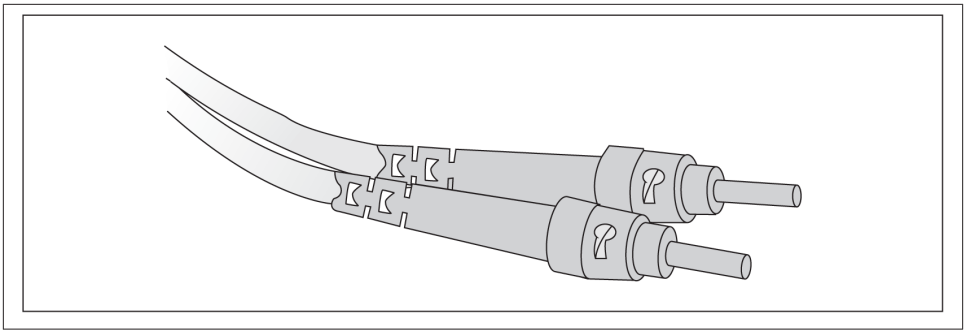


Figure 8-3. ST connectors

Connecting a 10BASE-FL Ethernet Segment

The 10BASE-FL full-duplex segment can be used as a link segment between Ethernet switches equipped with 10BASE-FL ports, or between two media converters.

Figure 8-4 shows a 10BASE-FL link between two media converters. The 10BASE-FL port on the media converter is connected over a fiber optic cable to another 10BASE-FL port on a converter at the other end of the link. A signal crossover is required to make a connection between the two 10BASE-FL ports.

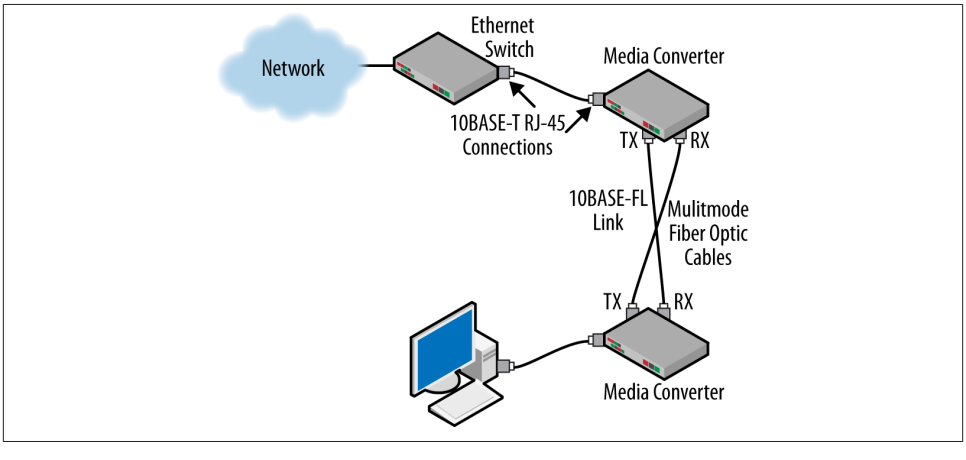


Figure 8-4. Connecting a 10BASE-FL Ethernet link

10BASE-FL Link Integrity Test

10BASE-FL transceivers monitor the light level on the fiber optic link segment to provide a link integrity test. Vendors can optionally provide a link light on the fiber optic transceiver to give you a visual indication of the link's integrity status. If the link lights

on the transceivers at each end of the link are lit when you connect them to the segment, then you know that both transceivers are powered up and working. The lights also indicate that the segment is connected properly, and that the optical loss is within acceptable limits.

To provide continual link detection, the 10BASE-Fl transceivers send a 1 MHz idle signal during periods when no data is being sent. If the light level on the link drops below that required for reliable data reception, the transceivers will detect this condition and stop sending or receiving data over the link. However, transmission of the idle signal will continue, which provides a way to detect the link when the light level over the fiber optic link returns to an acceptable value.

10BASE-FL Configuration Guidelines

The Ethernet standard contains guidelines for a single 10BASE-FL fiber optic segment. These guidelines are listed in Table 8-3.

Table 8-3. 10BASE-FL single segment guidelines

Media type	Maximum segment length	Maximum number of transceivers (per segment)
10BASE-FL	2,000 m (6,561 feet)	2

There is no minimum length specified for this segment type. However, some vendors have offered "extended length" versions of this equipment that require a vendor-specific minimum-length segment to prevent signal errors caused by over-driving the fiber optic receiver.

Longer 10 Mb/s fiber segments

Longer fiber segments are possible when the link is operated in full-duplex mode. The use of full-duplex mode on a link segment means that the segment length is no longer restricted by the round-trip timing limits of a shared Ethernet channel. Instead, the segment length is limited only by the signal-carrying characteristics of the media; in this case, the optical power loss (signal attenuation) and signal dispersion over the fiber optic cable. Vendors have produced transceivers that can achieve distances of up to 5 kilometers (km) over full-duplex segments built using multimode fiber optic cables.

Single-mode fiber optic cable transceivers can be purchased to drive a full-duplex 10 Mb/s link for distances of up to 40 km. However, a single-mode fiber system is more expensive and difficult to use than multimode. The single-mode fiber optic core is typically 8 or 9 μm in diameter, compared to the 62.5 μm core in multimode cable. Coupling a light source into the small core of single-mode cable requires a more expensive laser light source and precise connectors. Therefore, while much longer full-duplex fiber optic segments are possible, you must be prepared to deal with more complex fiber optic design and installation issues.

Today, longer fiber optic link segments typically operate at higher speeds because of the improved throughput that the higher speeds provide. Given that fiber optic links are often installed as backbone links and uplinks between switches, they are normally operated at the highest speed provided by the Ethernet interfaces at each end. Most Ethernet switches provide 100 Mb/s or 1 Gb/s fiber optic uplink speeds, and uplinks supporting 10 gigabits/s have become more common. Uplink ports that support 40 Gb/s are now available and are being adopted as network throughput requirements continue to increase.

CHAPTER 9

100 Mb/s Ethernet

This chapter describes the signaling and media components used in the 100BASE-TX and 100BASE-FX systems. The 100 Megabit per second "Fast Ethernet" media systems were first defined in the 802.3u supplement to the Ethernet standard in 1995. These systems are still in wide use, providing high-speed service at low cost to desktop computers and other devices.

The most widely used 100 Mb/s media standards are based on specifications first developed in the 1990s for the Fiber Distributed Data Interface (FDDI) network standard. After the development of 100 Mb/s Ethernet technology, equipment based on the FDDI standard lost market share and is no longer sold, but FDDI technology lives on in the 100BASE-X Ethernet standards, which include twisted-pair and fiber optic cable types.

100BASE-X Media Systems

The 100BASE-X system includes 100BASE-TX twisted-pair and 100BASE-FX fiber optic segments based on FDDI technology. Although multiple 100 Mb/s copper media systems were developed, the 100BASE-X media segments became the most widely adopted, and the other systems are no longer sold.

The systems that are no longer sold include:

100BASE-T4
: Designed to use four pairs of Category 3 or better twisted-pair cabling.

100BASE-T2
: Designed to use two pairs of Category 3 or better cabling.

Fast Ethernet Twisted-Pair Media Systems (100BASE-TX)

The 100BASE-TX twisted-pair media system is based on the ANSI FDDI TP-PMD (Twisted-Pair Physical Medium Dependent) standard. The system operates over two pairs of twisted-pair wires: one pair to receive data signals, and the other pair to transmit data signals.

100BASE-TX Signaling Components

The following signaling components in the 100BASE-TX system may be used to send and receive signals over a twisted-pair cable segment:

- An Ethernet interface with a built-in 100BASE-TX transceiver.
- A 100BASE-TX transceiver, also called a *physical layer device* (PHY).
- A medium-independent interface (MII). Externally exposed MIIs are no longer used on Ethernet interfaces. More details on the external Fast Ethernet MII may be found in Appendix C.

Today, all 100BASE-TX connections are made to RJ45 connectors on built-in Ethernet interfaces in computers and switch ports that are equipped with built-in transceivers.

100BASE-TX Ethernet Interface

Modern 100BASE-TX interfaces are equipped with a built-in 100BASE-TX transceiver used to make a direct connection to the twisted-pair segment.

Figure 9-1 shows a connection between a desktop computer and an Ethernet switch port over a 100BASE-TX twisted-pair segment. The computer is equipped with an Ethernet interface that supports 100BASE-TX operation. The interface comes with an RJ45-style jack connector, to which the RJ45 plug on the end of the twisted-pair cable is connected.

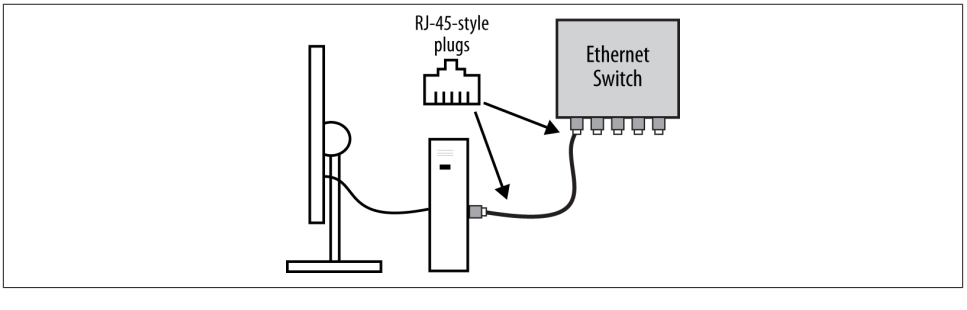


Figure 9-1. 100BASE-TX Ethernet interface

The Ethernet switch is shown with RJ45 connectors connecting to the built-in Ethernet interfaces in each switch port. Automatic signal crossover is usually performed by the interface electronics, so that the cable segment and patch cable can be wired "straight through."

Twisted-pair Ethernet interfaces on commonly available desktop computers and other devices often support three speeds—10, 100, and 1000 Mb/s—and use Auto-Negotiation to find the fastest common speed supported over a given connection. Lower-cost Ethernet switches used for desktop connections often support 10 and 100 Mb/s port speeds on the ports intended for station connections, resulting in 100 Mb/s operation when connected to a desktop computer interface over a twisted-pair cable segment.

Switches with 10, 100, and 1000 Mb/s support on all ports have become commonly available, and as prices have continued to drop, switches with this level of support for all ports are becoming more widely adopted. Switches that support these port speeds often have uplink ports operating at 1 or 10 Gb/s, as described in Chapter 19.

100BASE-TX Signal Encoding

The 100BASE-TX system is based on the signaling originally developed for the ANSI X3T9.5 FDDI standard, which includes both fiber optic and twisted-pair media. The signal encoding used on Fast Ethernet systems is based on block encoding, which is more complex than the original Manchester encoding system used in 10 Mb/s Ethernet. Block encoding takes a group, or block, of data bits and encodes them into a larger set of code bits.

The data stream is divided into a fixed number of bits per block—usually 4 or 8. Each data block is translated into a set of code bits, also called *code symbols*. For example, a 4-bit data block (16 possible bit patterns) may be translated into a 5-bit code symbol (32 possible values). The expanded set of code symbols is carefully chosen, and the bit patterns of the individual symbols are designed to help improve line signaling by providing a better balance between ones and zeros. The extra code symbols are used for control purposes, such as start-of-frame, end-of-frame, and error signaling.

Depending on the media system, the block-encoded symbols may be transmitted using a simple two-level signaling system, or more complex multilevel line signaling. This effectively compresses the number of bits being transmitted into a smaller number of signal transitions on the cable. Using more complex line signaling schemes results in signal transition rates that can be supported on typical twisted-pair cable, which has limits on how fast it can transmit data.

The 100BASE-TX and 100BASE-FX Fast Ethernet systems (collectively known as 100BASE-X) both use the same block encoding scheme. The 100BASE-T4 and 100BASE-T2 systems use different encoding schemes and physical line signaling to provide Fast Ethernet signaling over twisted-pair cable that is of lower quality and not

rated to meet Category 5 specifications. However, because Category 5 cabling is widely adopted, there is no need for these two systems, and they never caught on in the marketplace; therefore, they will not be discussed here.

100BASE-X encoding

Rather than reinventing the wheel when it came to sending signals at 100 Mb/s, the 100BASE-X media systems adopted portions of the block encoding and physical line signaling originally developed for the ANSI Fiber Distributed Data Interface (FDDI) network standard. FDDI is a 100 Mb/s Token Ring network that was popular in the early 1990s. The block encoding used in FDDI and the 100BASE-X media types relies on a system called *4B/5B*, which divides the data into 4-bit blocks. The 4-bit blocks are translated into 5-bit code symbols for transmission over the media system. The encoded symbols are transmitted over fiber optic cables as two-level signals. The addition of a fifth bit means that the 100 Mb/s data stream becomes a 125-Mbaud stream of signals on a fiber optic media system.

A *baud* is a unit of signaling speed per second. One Megabaud (Mbaud) = one million baud, or one million signaling events per second.

The 5-bit encoding scheme allows for the transmission of 32 distinct 5-bit symbols, including 16 symbols that carry the 4-bit data values from 0 through F (hexadecimal), along with a further set of 16 symbols used for control and other purposes. These other symbols include the IDLE symbol, which is continually sent when no other data is present. ("IDLE" is not an acronym. Instead, IDLE is in uppercase in the Ethernet standard to show that the word is formally defined.) The IDLE symbol is used to keep the signaling system active when there is no other data to send. For this reason, the signaling system used in Fast Ethernet is continually active, sending IDLE symbols at 125 Mbaud if nothing else is going on (unless Energy Efficient Ethernet is used to minimize signaling and save power).

Table 9-1 shows the entire code space of 5-bit symbols that can be sent over the channel. The 5-bit data symbols transmitted on the channel are mapped to 4-bit "nibbles" of data, which are sent over the MII interface. The Data 0 through F and the IDLE and SLEEP symbols are all considered data symbols. The symbols J, K, T, and R are considered control symbols, and are used for special purposes such as indicating the start of the frame. The rest of the symbols are unused and marked as Invalid in the standard.

Table 9-1. Five-bit symbols

Five-bit code group as sent on the channel	Name	Data on MII interface	Interpretation
---- 11110 ----	0	---- 0000 ----	Data 0
---- 01001 ----	1	---- 0001 ----	Data 1
---- 10100 ----	2	---- 0010 ----	Data 2
---- 10101 ----	3	---- 0011 ----	Data 3
---- 01010 ----	4	---- 0100 ----	Data 4
---- 01011 ----	5	---- 0101 ----	Data 5
---- 01110 ----	6	---- 0110 ----	Data 6
---- 01111 ----	7	---- 0111 ----	Data 7
---- 10010 ----	8	---- 1000 ----	Data 8
---- 10011 ----	9	---- 1001 ----	Data 9
---- 10110 ----	A	---- 1010 ----	Data A
---- 10111 ----	B	---- 1011 ----	Data B
---- 11010 ----	C	---- 1100 ----	Data C
---- 11011 ----	D	---- 1101 ----	Data D
---- 11100 ----	E	---- 1110 ----	Data E
---- 11101 ----	F	---- 1111 ----	Data F
---- 11111 ----	I	Undefined	IDLE; used as interstream fill code
---- 00000 ----	P	Undefined	SLEEP—LPI code only for EEE capability; otherwise, Invalid code
---- 11000 ----	J	---- 0101 ----	Start-of-Stream delimiter, Part 1 of 2; always used in pairs with K
---- 10001 ----	K	---- 0101 ----	Start-of-Stream delimiter, Part 2 of 2; always used in pairs with J
---- 01101 ----	T	Undefined	End-of-Stream Delimiter, Part 1 of 2; always used in pairs with R
---- 00111 ----	R	Undefined	End-of-Stream Delimiter, Part 2 of 2; always used in pairs with T
---- 00100 ----	H	Undefined	Transmit Error; used to force signaling errors
---- 00000 ----	V	Undefined	Invalid code
---- 00001 ----	V	Undefined	Invalid code
---- 00010 ----	V	Undefined	Invalid code
---- 00011 ----	V	Undefined	Invalid code
---- 00101 ----	V	Undefined	Invalid code
---- 00110 ----	V	Undefined	Invalid code
---- 01000 ----	V	Undefined	Invalid code

Five-bit code group as sent on the channel	Name	Data on MII interface	Interpretation
---- 01100 ----	V	Undefined	Invalid code
---- 10000 ----	V	Undefined	Invalid code
---- 11001 ----	V	Undefined	Invalid code

Carrier detection in MII transceivers only becomes active when actual frame data symbols are seen on the channel. The pair of symbols called J and K are used together to indicate the start of the preamble in an Ethernet frame. The pair of symbols called T and R are used to indicate the end of the frame. The PHY deals with the task of recognizing these 5-bit symbols, removing the special symbols, and delivering standard Ethernet frame data to the interface.

Each Fast Ethernet media system uses a different line signaling scheme to send the block-encoded signals over the physical media.

100BASE-TX physical line signaling

The physical signaling used to transmit 5-bit symbols over twisted-pair cables is based on a system called *multilevel threshold-3* (MLT-3). This means that during each signal period, the signal can be at one of three levels: +, 0, or –. During each bit time, a change from one level to the next indicates a logical one (1), whereas no change in signal level indicates a logical zero (0), as shown in Figure 9-2. Because the signal level doesn't change when a zero is transmitted, this reduces the total number of signal transitions on the wire.

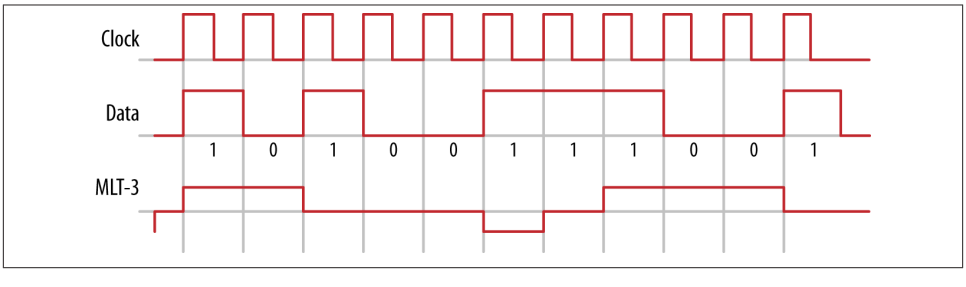


Figure 9-2. MLT3 signaling

In the 100BASE-TX transceiver, the 4B/5B block-encoded data is first scrambled using "pseudorandom" values that make it possible for the receiver to run the same algorithm and descramble the data. Scrambling is done to spread out the electromagnetic emission patterns in the data. By scrambling the data, the system avoids sending signals at a single frequency, which can otherwise increase interference.

A single frequency of signaling is caused by the transmission of repetitive data patterns, such as a continuous series of ones or zeros. The result of scrambling the data is that no single frequency is sent for any significant period of time, and the signaling power is spread out over a range in the frequency spectrum. This reduces interference and makes more efficient use of the available bandwidth.

The scrambled data is transmitted onto the twisted wire pairs as a series of three voltages with a signal transition rate of 125 Mbaud. The differential voltages swing from approximately zero to +1 volts on the positive wire, and from zero to –1 volts on the negative wire of the pair.

Even though the MLT-3 signaling system reduces the signaling rate, the 100BASE-TX system is still doing a lot of high-frequency signaling over the twisted-pair cables. Therefore, it's important that all twisted-pair cables, including patch cords and other components used in a 100BASE-TX segment, meet or exceed the Category 5 signal-carrying specifications needed to handle the signals. If lower-quality cables and components are used, the signal error rate will increase, causing frame loss and reduced network performance.

100BASE-TX Media Components

The following set of media components are used to build a 100BASE-TX twisted-pair segment:

- Unshielded or shielded twisted-pair cable
- Eight-position RJ45-style modular connectors that meet Category 5 specifications

UTP cable

The 100BASE-TX system operates over two pairs of unshielded twisted-pair (UTP) wires; one pair receives data signals, while the other pair transmits data signals. The maximum segment length is 100 meters (328.08 feet) of unshielded twisted-pair cable that has a 100 ohm characteristic impedance rating and that meets or exceeds the TIA/EIA Category 5 specifications. Additional details on installing and using UTP cables and connectors can be found in Chapter 15.

Eight-position RJ45-style jack connectors

The 100BASE-TX system requires two pairs of wires that are terminated in an eight-position (RJ45-style) connector, which means that four pins of the eight-position connector are used. The 100BASE-TX signals used on the 8-pin connector are the same as the 10BASE-T signals shown in Table 8-1.

The pin numbers used in the eight-pin connector for 100BASE-TX were changed from the ones defined in the FDDI TP-PMD standard in order to conform to the wiring

scheme already in use in the 10BASE-T standard. The ANSI standard uses pins 7 and 8 for receive data, whereas 100BASE-TX uses the same pins as the 10BASE-T system: 3 and 6. That way, an Ethernet interface supporting 10BASE-T and 100BASE-X can use the same Category 5 cabling system.

According to the structured cabling standards, a twisted-pair segment will have all eight wires connected to the RJ45-style connector, even though the 100BASE-TX media system only uses four of the eight wires. The other wires should not be used to support any other services, as the 100BASE-TX system is not designed to tolerate the increased signal crosstalk that occurs when sharing the cable with other signals. Connecting all eight wires makes it possible for the cable segment to support higher-speed Ethernet media systems that use all four signal pairs.

100BASE-TX Link Integrity Test

The 100BASE-TX transceiver circuits (PHY) continually monitor the receive data path for activity as a means of checking that the link is working correctly. The signal encoding system used on 100BASE-TX segments results in signals being sent continually, even during idle periods. Therefore, activity (i.e., the reception of IDLE symbols) on the receive data path is sufficient to provide a continual check of link integrity.

100BASE-TX Configuration Guidelines

Table 9-2 lists the single segment guidelines for a 100BASE-TX segment. The 100BASE-TX specifications allow a segment with a maximum length of 100 meters.

Table 9-2. 100BASE-TX single segment guidelines

Media type	Maximum segment length	Maximum number of interface connections (per segment)
Twisted-pair 100BASE-TX	100 m (328.08 feet)	2

There is no minimum length specification for a 100BASE-TX segment. In practice, you can purchase ready-made patch cables as short as one foot and use them to connect 100BASE-TX equipment together. However, you may find that if you want to test the cable with a handheld cable tester, there may be a minimum cable length that the tester requires for an accurate test of cable parameters.

Fast Ethernet Fiber Optic Media Systems (100BASE-FX)

The 100BASE-FX fiber optic media system provides all of the advantages of the older 10BASE-FL fiber optic link segment, while operating 10 times faster. Distances of 2 km (6561.6 feet) over multimode fiber optic cables are possible when operating 100BASE-FX segments in full-duplex mode. Considerably longer distances are possible when using single-mode fiber segments.

While 100BASE-FX segments were initially widely used for switch uplinks, newer standards operating at faster speeds are now most often used for that purpose. 100BASE-FX fiber segments are still in use, but new networks and network upgrade designs usually use 1 Gb/s or 10 Gb/s uplinks for the higher performance that they provide.

100BASE-FX Signaling Components

The following signaling components can be used in the 100BASE-FX system to send and receive signals:

- An Ethernet interface with a built-in 100BASE-FX fiber optic transceiver
- An external 100BASE-FX transceiver, also called a physical layer device (PHY)

We'll look at the transceivers shortly.

100BASE-FX Signal Encoding

The 100BASE-FX system is based on block-encoded signaling originally developed for the ANSI X3T9.5 FDDI standard, which includes both fiber optic and twisted-pair media. The block encoding used in FDDI and 100BASE-FX relies on a system called 4B/5B, described earlier in this chapter (see "100BASE-X encoding" on page 142).

Physical line signaling

The physical signaling used to transmit 100BASE-FX signals is accomplished by sending light pulses over the fiber optic cables. The 100BASE-FX system uses a variant of the Non-Return-to-Zero (NRZ) scheme called Non-Return-to-Zero Inverted (NRZI).

This system makes no change in the signal level when sending a logical zero (0), and inverts the signal from its previous state for a logical one (1). The goal is to ensure a minimum number of logic transitions in the signal to provide clocking information for the signal decoding circuits.

The peak optical transmission power from a 100BASE-FX transceiver is between 200 and 400 microwatts (μW). Given an approximately equal number of ones and zeros sent over the segment, the average power sent over a fiber optic link is between 100 and 200 μW. These figures are for light being coupled into a 62.5/125 micron (μm) fiber, rated as OM1. Because there are no extraneous electromagnetic emissions on a fiber optic link, there is no need to scramble the data, as done with 100BASE-TX systems.

100BASE-FX Media Components

The following set of media components are used to build a 100BASE-FX fiber optic segment:

- Fiber optic cable
- Fiber optic connectors

Fiber optic cable

The 100BASE-FX specification requires two strands of *multimode fiber optic* (MMF) cable per link, one for transmit data and one for receive data, with the signal crossover (TX to RX) performed in the link. There are many kinds of fiber optic cables available, ranging from simple two-strand jumper cables with PVC plastic for the outer jacket material on up to large interbuilding cables carrying many fibers in a bundle.

The minimum-quality fiber optic cable used for a 100BASE-FX fiber link segment is classified as OM1 according to the ISO 11801 standard, and consists of a *graded-index* MMF cable. This cable has a 62.5 µm fiber optic core and 125 µm outer cladding (62.5/125). The light wavelength used on a 100BASE-FX fiber link segment is 1,350 nanometers (nm). Signals sent at that wavelength over MMF fiber can provide segment lengths of up to 2,000 meters (6,561 feet) when operating the link in full-duplex mode. More details on fiber optic cables and connectors can be found in Chapter 17.

Fiber optic connectors

According to the original standard, the medium dependent interface (MDI) for a 100BASE-FX link was defined as one of three kinds of fiber optic connector. Of the three, the duplex SC connector, shown in Figure 9-3, was the recommended alternative in the standard, and it was widely used by vendors. The SC connector is designed for ease of use; the connector is pushed into place and automatically snaps into the connector housing to complete the connection.

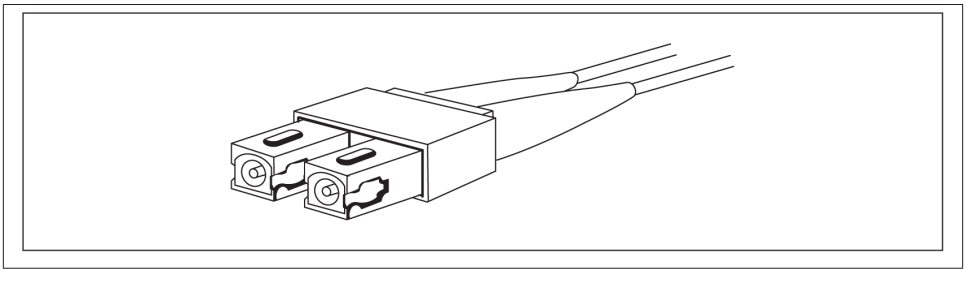


Figure 9-3. Duplex SC fiber optic plug

When the 100BASE-FX standard was first adopted, vendors supported the SC fiber optic connector as described in the standard. Over time, the number of fiber optic standards increased, and a market developed for switches that could support multiple types of fiber optic connections.

In response, vendors developed a *small form-factor pluggable* (SFP) transceiver, which can be purchased to support several different kinds of Ethernet fiber optic media systems. The SFP transceiver uses a smaller fiber optic connector called the LC connector.

Figure 9-4 shows the smaller LC fiber optic plug, which is used for connections to SFP fiber optic transceivers.

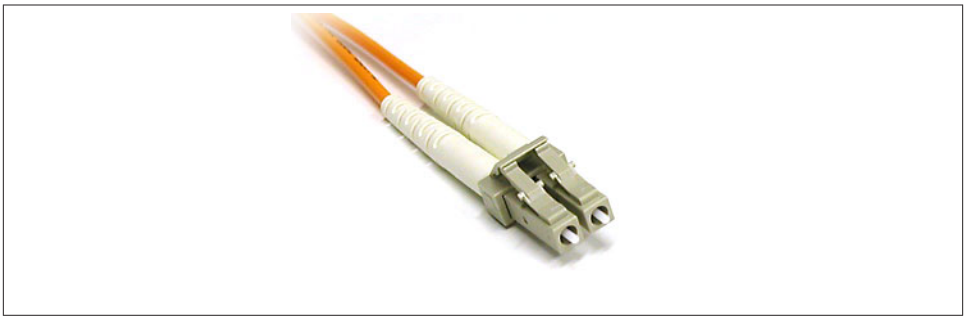


Figure 9-4. Duplex LC fiber optic plug

100BASE-FX transceivers

A 100BASE-FX Ethernet interface with a built-in transceiver is connected directly to a fiber optic Ethernet segment; there is no need for an outboard transceiver, because the transceiver is built into the interface card.

Figure 9-5 shows an SFP transceiver. The 100BASE-FX SFP transceiver module is plugged into the switch port or network interface card, and a fiber optic cable equipped with an LC plug is connected to the LC fiber optic socket on the end of the SFP transceiver.

Figure 9-5. 100BASE-FX SFP transceiver

100BASE-FX Fiber Optic Characteristics

There is an 11 dB optical loss budget allowed per 100BASE-FX segment. A typical performance rating for OM1 multimode fiber operating at 1,350 nm provides roughly 1 dB loss per 1,000 m of cable. You can also expect something in the neighborhood of 0.5 to 1.5 dB loss per connection, depending on how well the connection has been made.

Alternate 100BASE-FX Fiber Optic Cables

The ANSI media standard, upon which 100BASE-FX is based, notes that alternate multimode fiber optic cables may be used. This includes cables with a 50 µm fiber optic core and 125 µm outer cladding (50/125), cables with an 85 µm core and 125 µm cladding (85/125), and cables with a 100 µm core and 125 µm cladding (100/125). The difficulty is the same as with the 10BASE-FL system: the mismatch between the size of the fiber optic core on the alternative cables and the 62.5 µm size of the receiver and transmitters on 100BASE-FX equipment.

The same issues apply to alternative cables for 100BASE-FX as for 10BASE-FL (discussed in Chapter 8): there will be a significant loss in signal due to the mismatch in core size. To compensate for the losses, the total length of the segment must be reduced.

100BASE-FX Link Integrity Test

The 100BASE-FX transceiver circuits (PHY) continually monitor the receive data path for activity as a means of checking that the link is working correctly. The signaling system fires continually even during idle periods of no network traffic. Therefore, activity on the receive data path is sufficient to provide a continual check of link integrity.

100BASE-FX Configuration Guidelines

The Ethernet standard contains guidelines for building a single 100BASE-FX fiber optic segment. These are listed in Table 9-3.

Table 9-3. 100BASE-FX single segment guidelines

Media type	Maximum segment length	Maximum number of transceivers (per segment)
Fiber optic 100BASE-FX	2 km (6,561.68 feet)	2

The maximum segment length is for a full-duplex segment connected between two Ethernet transceivers over OM1 multimode fiber. There is no minimum length specified for this segment type. Two 100BASE-FX stations can be linked with a patch cable that is as short as practicable.

Long Fiber Segments

Long fiber segments are possible when the link is operated in full-duplex mode. In full-duplex mode, the segment length is no longer restricted by the round-trip timing limits of a shared Ethernet channel. Instead, the segment length is limited by the optical power loss (signal attenuation) and signal dispersion over the fiber optic cable. Typical fiber optic transceivers can achieve distances of 2 km over 100BASE-FX segments built using multimode fiber optic cables. Longer distances (from 40 km up to 80 km) can be achieved when using single-mode fiber for a full-duplex segment.

While single-mode 100BASE-FX links can achieve distances of 40 km or more, this type of fiber is more expensive and difficult to use than multimode fiber. The single-mode fiber core is usually 8 or 9 μm in diameter, compared to the 62.5 μm core in multimode fiber. Coupling a light source into the smaller core of single-mode fiber requires a more expensive laser light source and precise connectors.

CHAPTER 10

Gigabit Ethernet

The IEEE standard uses both "1000 Mb/s" and "Gigabit Ethernet" to describe this variety of Ethernet media system, which operates over both twisted-pair and fiber optic cabling.

The specifications for the 1000BASE-X system for fiber optic media were developed in the 802.3z supplement to the IEEE standard, which was adopted in 1998 as Clauses 34–39 of the standard. The specifications for the 1000BASE-T twisted-pair media system were developed in the 802.3ab supplement to the IEEE standard, which was adopted in 1999 as Clause 40 of the standard.

Gigabit Ethernet Twisted-Pair Media Systems (1000BASE-T)

Supporting 1 billion bits per second over unshielded twisted-pair (UTP) cable was a remarkable achievement at the time that this standard was developed. To make it happen, the 1000BASE-T media system uses a mix of signaling and encoding techniques that were originally developed for the 100BASE-TX, 100BASE-T2, and 100BASE-T4 media standards. While 100BASE-T2 and 100BASE-T4 were not widely adopted in the marketplace, their technology was essential to developing the 1000BASE-T standard.

The 100BASE-T2 Fast Ethernet standard is based on a signal encoding system that could send 100 Mb/s Ethernet signals over two pairs of Category 3 cable. These signaling techniques were adopted and extended by the 1000BASE-T standard for use over four pairs of twisted-pair cable rated at Category 5 or better.

From the 100BASE-T4 system, the 1000BASE-T standard adopted the technique of simultaneously sending and receiving signals over the same wire pairs. The 1000BASE-T system also adopted the line signaling rate of the very popular 100BASE-TX Fast Ethernet system. Maintaining the same line signaling rate makes it possible for

1000BASE-T to work over the same widely used Category 5/5e cabling that also supports a 100BASE-TX link.

1000BASE-T Signaling Components

The 1000BASE-T interface comes with a built-in transceiver used to make a direct connection to the 1000BASE-T twisted-pair segment. The interface electronics can either be built into the computer at the factory, or be on an adapter card that is installed in one of the computer's expansion slots.

Unlike the original 10 and 100 Mb/s Ethernet systems that could provide an exposed AUI or MII connector to support an external transceiver and transceiver cable, the 1000BASE-T Gigabit Ethernet system requires an Ethernet interface with a built-in Gigabit Ethernet transceiver. There is no exposed transceiver connector in the Gigabit Ethernet system, and therefore no support for an external transceiver for copper media.

The external transceivers used in the original 10 and 100 Mb/s systems are no longer sold as new equipment.

Figure 10-1 shows a desktop computer connected to a switch port that is capable of operating at 1 Gb/s. The internal network interfaces use a combination of transceiver electronics to support operation at multiple speeds.

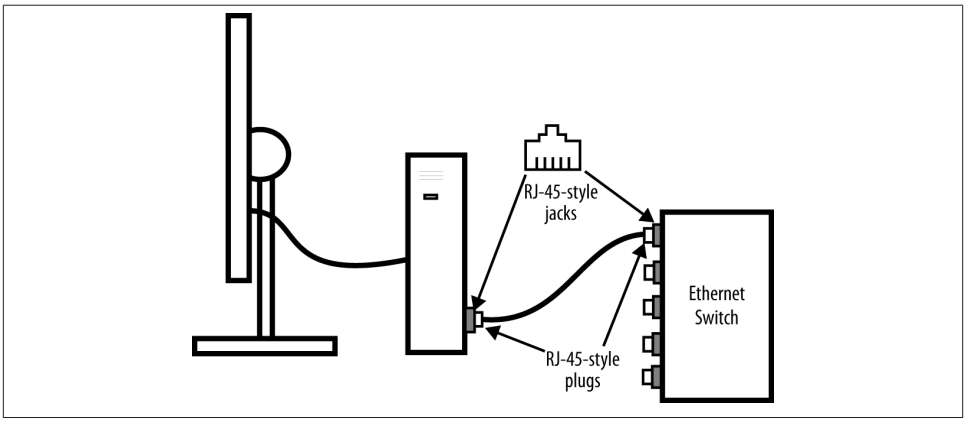


Figure 10-1. 1000BASE-T Ethernet interface

The Auto-Negotiation standard is typically used on multispeed interfaces to automatically configure the speed of operation over the link. The Ethernet switch has multiple

ports with built-in transceivers and Ethernet interfaces that are capable of operating at multiple media speeds.

1000BASE-T Signal Encoding

As mentioned earlier, signaling techniques that were originally developed for the 100BASE-T2, -T4, and -TX standards have been adopted and extended for Gigabit Ethernet. To this preexisting set of technologies, the 1000BASE-T system adds its own set of digital signaling processing techniques.

Signal encoding on a 1000BASE-T link is based on a block encoding scheme called 4D-PAM5, which refers to the combined four-dimensional trellis modulation and the five-level pulse amplitude modulation coding technique used. Trellis modulation is so called because a state diagram of the technique, when drawn on paper, resembles a trellis lattice used in gardening. The complete encoding scheme and the set of encoded symbols used are complex, and of primary interest only to designers of Ethernet interface chips.

The details of the encoding scheme and bit-to-symbol mapping are listed in Clause 40 of the 802.3-2012 Standard for Ethernet.

The five-level line signaling system includes forward error correction signals to improve the signal-to-noise ratio on the cable. The differential voltages used on the wire pairs swing from approximately +1 to –1 volt on both the positive and the negative wires.

Signaling and data rate

A 1000BASE-T link transmits and receives data on all four wire pairs simultaneously. The 1000BASE-T transceivers at each end of the link contain four identical transmit sections and four identical receive sections. Each of the four wire pairs in the link segment is connected to both transmit and receive circuitry in the transceiver. A circuit called a *hybrid* enables the transceiver to deal with the task of simultaneously transmitting and receiving signals on each wire pair.

Hybrid signaling has a long pedigree, having been used in the telephone industry to provide both transmit and receive signals on the single pair of wires connected to an analog telephone. The hybrid circuit for an analog telephone was designed to subtract the voice signal being transmitted from the received signal that is also on the line, so that you could hear the person you were talking with. A small amount of transmit signal was also directed to the earpiece, so that you could hear yourself as well.

Figure 10-2 is a schematic drawing of two transceivers connected together over four twisted pairs of wire. This drawing shows the basic data paths through the hybrid circuit, which provides simultaneous bidirectional transmission with echo cancellation. All four wire pairs are simultaneously used to send and receive data.

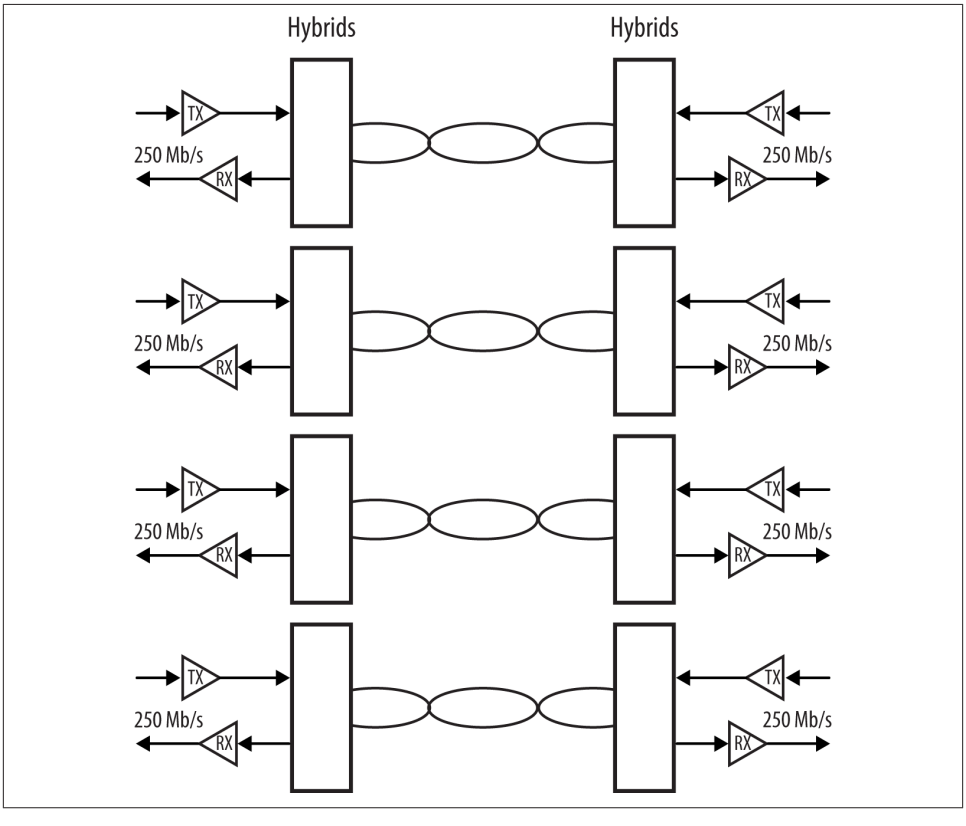


Figure 10-2. 1000BASE-T signal transmission

Two bits of Ethernet data are encoded and sent per signal transition on each wire pair. Thus, a total of eight bits of information is sent across all four pairs for each signal transition. A signal transition rate of 125 Mbaud therefore achieves a total data rate of 1000 Mb/s. Using a five-level line signaling system maintains approximately the same signaling rate on the cable as used by the 100BASE-TX Fast Ethernet system.

The continuous signaling in both directions on all four wire pairs generates signal echo and crosstalk, which the 1000BASE-T system must handle with a set of *digital signal processing* (DSP) techniques. These include echo cancellation, *near-end crosstalk* (NEXT) cancellation, and *far-end crosstalk* (FEXT) cancellation. Another DSP technique is *signal equalization*, to help compensate for signal distortion over the channel.

To spread out the electromagnetic emission patterns in the data to help avoid signal emissions from the cable, the 1000BASE-T transceiver also scrambles the signal.

Signal clocking

Auto-Negotiation support, mandatory in the 1000BASE-T standard, is embedded in the transceiver (PHY). To help improve signal processing, 1000BASE-T includes a *master-slave system* of synchronous signal clocking for each wire pair. The master and slave ends of a given wire pair synchronize their signaling using the clock signal provided by the master. This allows the transceiver circuits linked over each wire pair to differentiate between the signals they are sending and any other signals. This, in turn, makes it possible to strongly suppress signal echo, alien crosstalk, and far-end crosstalk signals, thus improving the signal-to-noise ratio. The Auto-Negotiation mechanism is used to decide which transceiver becomes the master and which the slave.

The signal encoding scheme provides both data symbols and symbols used for control and other purposes. These other symbols include the IDLE symbol, which is continually sent when no other data is present. The signaling system in 1000BASE-T is continually active; if nothing else is going on, it will send IDLE symbols at 125 Mbaud (unless Energy Efficient Ethernet is used to minimize signaling and save power).

1000BASE-T cabling requirements

A 1000BASE-T system operates at the same signaling rate over the cable as the 100BASE-TX system. However, the complex signaling techniques used in 1000BASE-T are more sensitive to certain signal performance issues on twisted-pair cable segments. Therefore, it's important that all twisted-pair cables and other components used in a 1000BASE-T segment meet or exceed the Category 5 signal-carrying specifications to properly handle the signals. The commonly used Category 5e cable has improved signal-carrying capabilities, and you can also use cables with even higher signal quality ratings, such as Category 6 and 6A.

Reliable Gigabit Ethernet operation requires that all patch cords be correctly assembled using high-quality components. The twisted pairs must maintain their twists as close as possible to the RJ45 connectors, and the connectors must be high quality for the best signal-carrying capabilities.

It can actually be quite difficult to build homemade patch cords that meet these requirements. Homemade patch cables that don't meet Category 5 specifications can cause problems on a 1000BASE-T segment. For the best results, you should purchase high-quality patch cords, manufactured under carefully controlled conditions, that are tested and rated to meet or exceed Category 5 specifications.

1000BASE-T Media Components

The following media components are used to build a 1000BASE-T twisted-pair segment:

- Category 5 UTP cable
- Eight-position RJ45-style modular connectors that meet or exceed Category 5 specifications

UTP cable

The 1000BASE-T system operates over four pairs of unshielded twisted-pair (UTP) wires. The maximum segment length is 100 meters (328.08 feet) of UTP cable that meets or exceeds the TIA/EIA Category 5/5e specifications. More details on installing and using twisted-pair cables and connectors can be found in Chapter 16.

Eight-position RJ45-style jack connectors

The 1000BASE-T media system uses four pairs of wires that are terminated in an eight-position (RJ45-style) connector. A 1000BASE-T system uses four pairs of wires, so all eight pins of the connector will be used.

As shown in Table 10-1, the four wire pairs are used to carry four bi-directional data signals (BI_D). The four signals are called BI_DA, BI_DB, BI_DC, and BI_DD. The data signals on each pair of a 1000BASE-T twisted-pair segment are polarized, with one wire of each signal pair carrying the positive (+) signal, and the other carrying the negative (–) signal. The signals are connected so that both wires associated with a given signal are members of a single wire pair.

Table 10-1. 1000BASE-T RJ45 signals

Pin number	Signal
1	---- BI_DA+ ----
2	---- BI_DA- ----
3	---- BI_DB+ ----
4	---- BI_DC+ ----
5	---- BI_DC- ----
6	---- BI_DB- ----
7	---- BI_DD+ ----
8	---- BI_DD- ----

The 1000BASE-T transceivers typically include circuits that can detect incorrect signal polarity (*polarity reversal*) in a wire pair. These circuits can correct polarity reversal by automatically moving the signals to the correct circuits inside the transceiver. However,

not all Ethernet devices may be able to correct a polarity reversal, so it is not a good idea to depend on this ability. Instead, all cables should be wired so that correct signal polarity is observed.

1000BASE-T Link Integrity Test

The Gigabit Ethernet transceiver circuits continually monitor the receive data path for activity as a means of verifying whether the link is working correctly. The signaling system used for 1000BASE-T segments continually sends signals—even during idle periods where there isn't any traffic on the network. Therefore, activity on the receive data path is sufficient to provide a check of link integrity.

1000BASE-T Configuration Guidelines

The Ethernet standard contains guidelines for building a single 1000BASE-T twisted-pair segment, shown in Table 10-2.

Table 10-2. 1000BASE-T single segment guidelines

Media type	Maximum segment length	Maximum number of transceivers (per segment)
Twisted-pair 1000BASE-T	100 m (328.08 feet)	2

There is no minimum length specification for a 1000BASE-T segment. In practice, you can purchase ready-made patch cables as short as one foot and use them to connect 1000BASE-T equipment together. However, you may find that if you want to test the cable with a handheld cable tester, there may be a minimum cable length that the tester requires for an accurate test of cable parameters.

The 1000BASE-T specification allows a segment with a maximum length of 100 meters. Unlike in a 10BASE-T system, 1000BASE-T segments cannot be longer than 100 meters due to signal transmission limits.

Gigabit Ethernet Fiber Optic Media Systems (1000BASE-X)

The 1000BASE-X identifier refers to three media segments: two fiber optic segments and a short copper jumper. Of these three, the fiber optic segments are widely used, while the short copper jumper was not adopted by the marketplace. Therefore, this chapter only describes the two fiber optic segments in detail.

The two fiber optic segments consist of a 1000BASE-SX (short wavelength) segment and a 1000BASE-LX (long wavelength) segment. The third segment type was called the 1000BASE-CX short copper jumper.

The 1000BASE-X media system is based on specifications first published in the ANSI X3T11 Fibre Channel standard. Fibre Channel is a high-speed network technology that was developed to support bulk data applications such as linking file servers for data

storage systems. The 1000BASE-X standard adapted the signal encoding and physical medium signaling from the Fibre Channel standard, with the only major change being an increase in the data rate from the 800 Mb/s rate used in Fibre Channel to 1000 Mb/s.

The first part of this section provides a brief look at 1000BASE-X signaling components and signal encoding, followed by a more detailed look at the 1000BASE-X media components.

1000BASE-X Signaling Components

The most widely available 1000BASE-X interfaces are designed for connection to a 1000BASE-SX media segment and support full-duplex mode only. The 1000BASE-SX system uses less-expensive short-distance lasers designed for connection to relatively short lengths of multimode fiber optic segments. Therefore, the 1000BASE-SX system is usually used inside buildings, and for connections to high-performance servers and workstations. The lengths of the fiber segments that each media system can accommodate are described later in this chapter.

A 1000BASE-X interface in an Ethernet switch port may support either 1000BASE-SX or 1000BASE-LX segments. High-performance switches typically support both 1000BASE-SX and 1000BASE-LX media types, as that results in the maximum flexibility. Smaller switches intended for use inside a single building may be equipped with only a fixed set of 1000BASE-SX ports.

The 1000BASE-CX short copper jumper was included in the standard for connections such as those inside a single machine room. However, it was not adopted by the marketplace, and 1000BASE-CX equipment was never made available.

A signal crossover is required to make the data flow correctly between two 1000BASE-X Ethernet interfaces. Signal crossover in fiber optic media systems is described in Chapter 17.

1000BASE-X Link Integrity Test

The Gigabit Ethernet transceiver circuits continually monitor the receive data path for activity to verify whether the link is working correctly or not. The signaling system used for 1000BASE-X segments sends signals continually, even during idle periods of no network traffic. Therefore, activity on the receive data path is sufficient to provide a check of link integrity.

1000BASE-X Signal Encoding

The 1000BASE-X system is based on signaling originally developed for the Fibre Channel standard. The Fibre Channel standard defines five layers of operation (FC0 through

FC4). The FC0 and FC1 layers are the ones adapted for use in Gigabit Ethernet. FC0 defines the basic physical link, including media interfaces that operate at various bit rates; FC1 defines signal encoding and decoding as well as error detection.

The block encoding used in Fibre Channel and 1000BASE-X is called *8B/10B*. In this encoding system, 8-bit bytes of data are turned into 10-bit code groups for transmission over the media system. The 10-bit encoding scheme allows for the transmission of 1,024 10-bit code groups. There are 256 code groups that carry the 8-bit data sent over the link, and another set of code groups that are used for special control characters.

The set of 1,024 10-bit code groups makes it possible to choose a specific set of 256 data code groups that contain sufficient signal transitions to ensure adequate clock recovery at the receiver end of the link. The code groups used to send data also ensure that the number of ones and zeros sent over time are approximately equal. This helps prevent any cumulative signal bias in the electronic components along the signal path that might otherwise be caused by the transmission of long strings of ones or zeros.

Special code groups are used for encoding the IDLE signal, which is continually sent when no other data is present, and to send signals that define the start and end of a frame. The complete set of data code groups and special code groups used is of primary interest only to designers of transceiver chips. Anyone who wants to examine the complete set of code groups can find them listed in Clause 36 of the Ethernet standard.

Physical line signaling

The physical signaling used to transmit the 10-bit code groups is based on the basic *Non-Return-to-Zero* (NRZ) line code. This is a simple line signaling code in which a logical one (1) results in a high voltage level or high light level, and a logical zero (0) results in a low voltage level or low light level.

Using 10 bits to encode every 8-bit byte and transmitting the signals with an NRZ line code causes the 1000 Mb/s Gigabit Ethernet data rate to become a 1,250,000 baud rate for signals on the media system. Because the maximum frequency at which light emitting diodes (LEDs) can operate is about 600 MHz, 1000BASE-X fiber optic transceivers must use lasers to handle the high-frequency signals.

1000BASE-X Media Components

The following set of media components are used to build a 1000BASE-X fiber optic segment:

- Fiber optic cable
- Fiber optic connectors

Gigabit Ethernet fiber optic segments use pulses of laser light instead of electrical currents to send Ethernet signals. This approach has several advantages. For one thing, a fiber optic link segment can carry Gigabit Ethernet signals for considerably longer distances than twisted-pair media can. The standard specifies that a full-duplex 1000BASE-LX segment must be able to reach as far as 5,000 meters (16,404 feet, or a little over 3 miles). However, most vendors sell "long haul" versions of 1000BASE-LX equipment that are designed to reach as far as 10 km (6.2 miles) on single-mode fiber. Vendors have also developed "extended reach" versions of 1000BASE-LX single-mode interfaces that can send signals over distances of 70–100 kilometers or more.

In large, multibuilding campuses, the fiber distances can add up fast, as the fiber cables may not be able to take the most direct route between buildings on the campus and a central switching location. Therefore, these long-reach transceivers can be quite useful. The LX interfaces are essential when it comes to building metropolitan area network (MAN) links, in which Gigabit Ethernet is used to provide network services between sites on a city-wide basis.

Fiber optic cable

Both 1000BASE-SX and 1000BASE-LX fiber optic media segments require two strands of cable: one for transmitting and one for receiving data. The required signal crossover, in which the transmit signal (TX) at one end is connected to the receive signal (RX) at the other end, is performed in the fiber optic link.

Maximum segment lengths for 1000BASE-SX and 1000BASE-LX are dependent on a number of factors. Fiber optic segment lengths in the Gigabit Ethernet system will vary depending on the cable type and wavelength used. More information on multimode and single-mode fiber optic segments and components can be found in Chapter 17.

Fiber optic connectors

The original standard recommended the use of duplex SC fiber optic connectors for both 1000BASE-SX and 1000BASE-LX fiber optic media segments. Figure 10-3 shows a duplex SC connector. Although the standard can recommend a connector, vendors can use other fiber optic connectors as long as they are not forbidden in the standard. For example, when the 1000BASE-X media systems first became available, vendors used the compact MT-RJ connector on 1000BASE-SX ports.

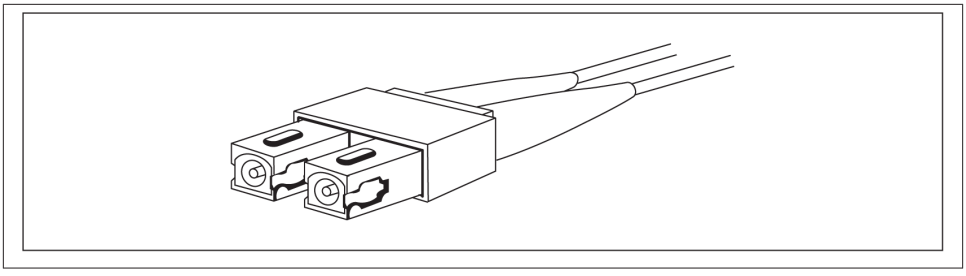


Figure 10-3. Duplex SC connector

Figure 10-4 shows the MT-RJ connector, which provided both fiber connections in a space the size of an RJ45 connector. Because the MT-RJ connector takes up about half the space required by the SC connectors, this allowed vendors to provide more 1000BASE-SX ports on switch.

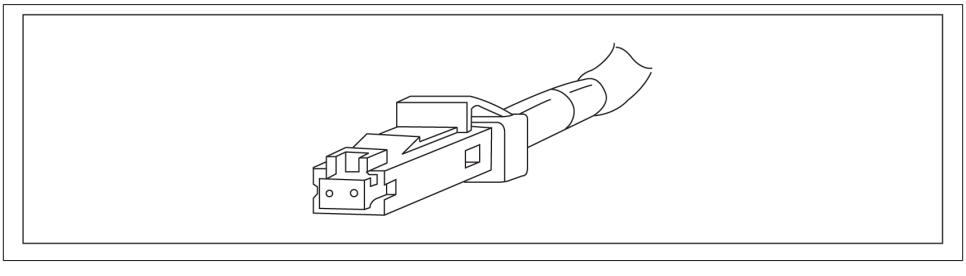


Figure 10-4. MT-RJ connector

1000BASE-X transceivers

Some vendors used the Gigabit Interface Converter (GBIC), which was an earlier form of transceiver module that allowed the customer to support either the 1000BASE-SX or 1000BASE-LX media types on a single port. The GBIC is a small, hot-swappable module that provides the media system signaling components for a Gigabit Ethernet port.

More recently, vendors have developed a *small form-factor pluggable* (SFP) transceiver, which can be purchased to support several different kinds of Ethernet fiber optic media systems.

The SFP transceiver is a small module that plugs into a switch port and uses a small fiber optic connector called the LC connector. Figure 10-5 shows the smaller LC fiber optic plug, which is used for connections to SFP fiber optic transceivers.

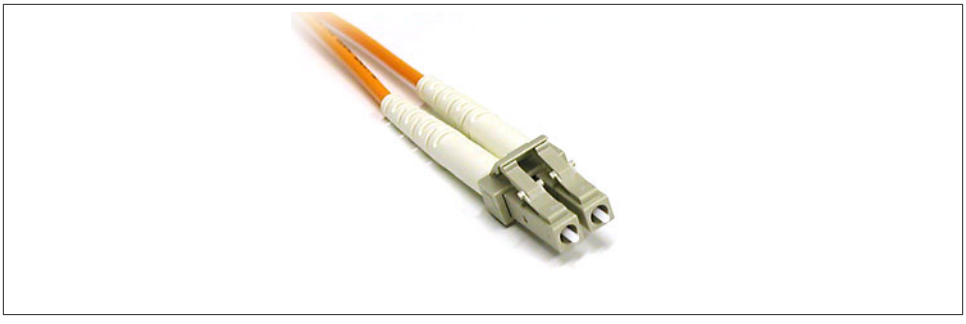


Figure 10-5. Duplex LC fiber optic plug

1000BASE-X Fiber Optic Specifications

The 1000BASE-SX short-wavelength media type operates at a wavelength of approximately 850 nm (770–860 nm is the range allowed in the specification) and requires multimode fiber optic cable. The 1000BASE-LX long-wavelength media type operates at a wavelength of approximately 1,300 nm (1,270–1,355 nm is the allowed range) and can be used with either multimode or single-mode fiber optic cable. You cannot see the laser light, as visible light ranges in wavelength from 455 nm (violet) to 750 nm (red). The 850 nm and longer wavelengths are in the infrared range.

In the Gigabit Ethernet system, the fiber optic loss budget is a major determinant of segment length. This section presents the worst-case optical loss budgets for the 1000BASE-SX and 1000BASE-LX segments, as well as the 1000BASE-LX/LH "long haul" segment. The 1000BASE-SX and 1000BASE-LX numbers are taken directly from the Gigabit Ethernet standard, and provide the typical maximum distance for the segments. According to the standard, the *minimum* distance between two stations for all segment types is 2.0 m (6.56 feet). Therefore, the shortest fiber optic cable you can use on any Gigabit Ethernet fiber link is 2 m.

In the following tables, "Channel insertion loss" accounts for the static power losses in the fiber optic cable, fiber optic jumper cables, and all connectors. The maximum distances in the tables are estimates based on the assumption that the total loss from connectors and splices will be 1.5 dB on a multimode link and 2.0 dB on a single-mode link.

1000BASE-SX Loss Budget

The 1000BASE-SX short-wavelength media type can only be used with multimode fiber. The distances achieved over multimode fiber vary according to the fiber specs.

Estimates from 2007 concluded that by 2010 more than 40% of the installed base of fiber optic cable in the United States would consist of OM1 and FDDI-grade fiber.[1]

The widely used TIA-568-A structured cabling standard also specified 160 MHz-km bandwidth for 62.5 µm cable at 850 nm, and 500 MHz-km at 1,300 nm.[2] Therefore, in the United States, the odds are good that OM1 62.5 µm MMF cabling installed prior to 1999 will have these ratings. Newer installations usually choose to use the more recent versions of multimode fiber, with multimode installations over the past several years usually based on OM3 cabling. The newest building cabling systems and data center designs often use the best cable that is available, which is currently OM4.

As shown in Table 10-3, the older OM1 62.5 µm multimode fiber with 160 MHz-km bandwidth only supported a 220 m maximum link distance. However, the new versions of OM3 and OM4 multimode fiber optimized for laser light make it possible to reach longer distances.

Table 10-3. Worst-case 1000BASE-SX loss budget and penalties

Parameter	62.5 µm MMF	62.5 µm MMF	50 µm MMF	50 µm MMF	Unit
Bandwidth measured at 850 nm wavelength	OM1 160	OM1 200	OM2 400	OM2 500	MHz-km
Link power budget	7.5	7.5	7.5	7.5	dB
Operating distance	220 721.78	275 902.23	500 1,640.42	550 1,804.46	meters feet
Channel insertion loss[a]	2.38	2.60	3.37	3.56	dB
Link power penalties[b]	4.27	4.29	4.07	3.57	dB
Unallocated margin	0.84	0.60	0.05	0.37	dB

[a] Operating distances used to calculate channel insertion loss are the maximum values.

[b] Link penalties are used for link budget calculations. They are not requirements and are not meant to be tested.

Table 10-4 provides some data on what happens when 1000BASE-X optics are connected to OM3 and OM4 cabling systems. The specific distances and loss budgets supported for this media type over these cables is not part of the original standard. Instead, you must consult the documentation for the transceivers used.

Table 10-4. 1000BASE-SX distance and loss budget on LOMF

Parameter	OM3 50 µm MMF	OM4 50 µm MMF	Unit
Channel insertion loss[a]	4.5	4.8	dB

1. See the Cisco "10GBASE-LRM and EDC: Enabling 10GB Deployment in the Enterprise" white paper.
2. The bandwidth for multimode fiber optic cable is specified as the product of megahertz times kilometers, shown as either MHz-km or MHz*km.

Parameter	OM3 50 µm MMF	OM4 50 µm MMF	Unit
Operating distance	550 1,804.46	550 1,804.46	meters feet[b]

[a] Operating distances used to calculate channel insertion loss are the maximum values.

[b] Longer lengths may be possible, per vendor documentation for the optical transceiver used.

1000BASE-LX Loss Budget

The 1000BASE-LX media type may be coupled to either multimode or single-mode fiber. When used with single-mode fiber, there is no modal dispersion and no differential mode delay effects, and the channel losses are quite a bit lower. For that reason, the distance achievable on single-mode fiber is much longer than with multimode fiber. When used with OM1 and OM2 multimode fiber, differential mode delay effects require that a mode-conditioning patch cable be installed for links over 300 m in length. This patch cable is described later in this chapter.

As you can see in Table 10-5, the longer wavelength (1,300 nm) used in 1000BASE-LX equipment also travels longer distances over typical multimode fibers. Given these longer distances, why not use 1000BASE-LX equipment everywhere? The answer has to do with cost. The lasers used in 1000BASE-LX equipment are two or three times more expensive than those used in 1000BASE-SX equipment.

Table 10-5. Worst-case 1000BASE-LX loss budget and penalties

Parameter	62.5 µm MMF	50 µm MMF	50 µm MMF	10 µm SMF	Unit
Bandwidth measured at 1,300 nm wavelength	500	400	500	N/A	MHz-km
Link power budget	7.5	7.5	7.5	8.0	dB
Operating distance	550 1,804.46	550 1,804.46	550 804.46	5,000 16,404.2	meters feet
Channel insertion loss[a]	2.35	2.35	2.35	4.57	dB
Link power penalties[b]	3.48	5.08	3.96	3.27	dB
Unallocated margin	1.67	0.07	1.19	0.16	dB

[a] Operating distances used to calculate channel insertion loss are the maximum values.

[b] Link penalties are used for link budget calculations. They are not requirements and are not meant to be tested.

1000BASE-LX/LH Long Haul Loss Budget

A widely used variant of the 1000BASE-LX media type includes a long haul (LH) transceiver that provides a more powerful laser. The higher-output laser makes it possible for Gigabit Ethernet signals to travel much longer distances over single-mode fiber. The details of the link power budget for long haul transceivers may vary, depending on the power of the transceiver. Therefore, you'll need to check with your vendor for details.

Table 10-6 contains the long haul power budget listed by one major vendor of equipment with 1000BASE-LX/LH ports, Cisco Systems. This particular long haul port type is based on a type of Gigabit Interface Converter (GBIC) that is widely used by other vendors as well.

Table 10-6. 1000BASE-LX/LH long haul loss budget

Parameter	10 µm SMF	Unit	Link power budget	10.5	dB
Operating distance	10,000 32,808.4	meters feet	Channel insertion loss[a]	7.8	dB
Link power penalties[b]	2.5	dB	Unallocated margin	0.2	dB

[a] Operating distances used to calculate channel insertion loss are the maximum values.

[b] Link penalties are used for link budget calculations. They are not requirements and are not meant to be tested.

1000BASE-SX and 1000BASE-LX Configuration Guidelines

The Ethernet standard contains guidelines for building a single 100BASE-FX fiber optic segment. Table 10-7 lists the single segment guidelines from the Ethernet standard for 1000BASE-SX and LX segments.

Table 10-7. 1000BASE-SX and 1000BASE-LX single segment guidelines

Media type	Minimum segment length	Maximum segment length	Maximum number of transceivers (per segment)
1000BASE-SX	2 m (6.5 feet)	220 m (721.78 feet)	2
1000BASE-LX	2 m (6.5 feet)	5,000 m (16,404.2 feet)	2

This segment length is for a full-duplex segment connected between two Ethernet transceivers over OM1 multimode fiber. In practice, transceiver vendors were able to provide longer distances, making it possible to build longer full-duplex segments. Check the documentation for the transceivers supported by the Ethernet switch port or network interface in question.

Differential Mode Delay

Differential mode delay (DMD) only occurs when a laser light source is connected to OM1 and OM2 multimode fiber. The DMD effect is a result of beam splitting caused by the way some OM1/OM2 multimode fiber cores were structured during manufacturing. This caused a small drop in the index of refraction, or *response curve*, of the cable. When a laser is coupled into the direct center of such a cable, two or more modes, or paths, can be excited. The multiple paths cause signals to arrive at the receiver at slightly different times, resulting in signal jitter. This, in turn, can make it difficult to demodulate the signal at the far end.

DMD does not occur in all OM1/OM2 fibers. Even in fibers that exhibit the problem, the amount of DMD may vary from one fiber to the next. Unfortunately, there is no reasonable way to field-test multimode cables for the presence of DMD. Carefully controlling the manufacture of multimode cables can entirely avoid DMD effects, but that is no help for existing fiber installations.

For incoherent light sources (LEDs) this is a nonissue, because all modes are simultaneously being used, which swamps any DMD effect. As it happens, this is also a nonissue for 1000BASE-SX links (despite their use of lasers), as the *coupled power ratio* over an SX link was high enough to swamp any signaling effects caused by DMD. In short, at the SX wavelength and link distances, the signal jitter caused by DMD is not a significant problem. However, DMD is still a major issue for 1000BASE-LX lasers when they are coupled to older OM1/OM2 multimode fiber.

Mode-Conditioning Patch Cord

The engineers on the standards committee found that DMD can be prevented in 1000BASE-LX links connected to OM1/OM2 multimode fiber cables by slightly offsetting the coupling of laser light into the cable. This avoids the beam splitting that can occur in some MMF cables when laser light is launched into the direct center of the cable. This type of offset signal launch is called *mode conditioning*. An outboard mode-conditioning patch cord must be used when an LX port is connected to a MMF fiber link.

Figure 10-6 shows the construction of a mode-conditioning patch cord. The cord contains a splice in the middle, in which the single-mode fiber is carefully connected to the multimode fiber with a slight offset from center. This keeps the single-mode laser light from entering the multimode cable at dead center, avoiding any DMD problems.

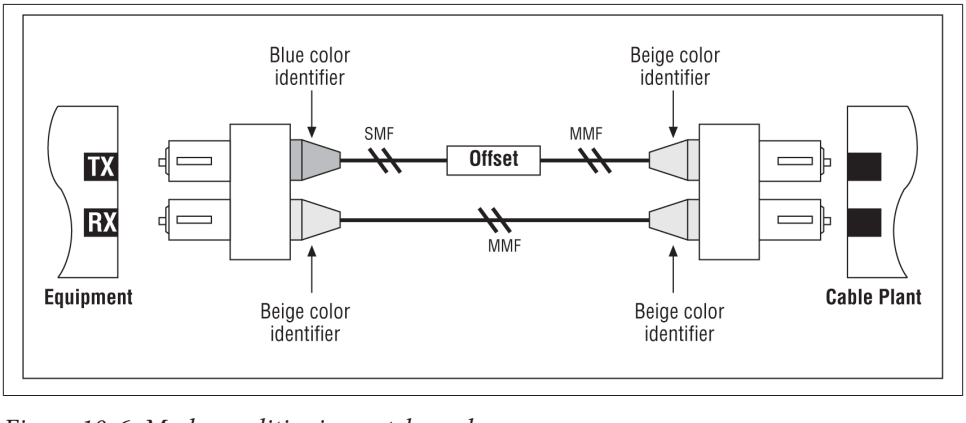


Figure 10-6. Mode conditioning patch cord

The standard notes that the single-mode end of the mode-conditioning patch cable should be labeled "To Equipment" and the multimode end "To Cable Plant." To help with identification, the plastic covering of the single-mode fiber connector should be blue and the multimode connectors should all be beige.

A mode-conditioning patch cord should be used at each end of the link when connecting 1000BASE-LX equipment to a multimode fiber optic segment. Make sure that the TX port on the equipment is connected to the single-mode portion of the mode-conditioning patch cable. The multimode fiber used in the conditioned launch cable should match the multimode fiber plant. In other words, if your fiber plant uses OM1 62.5/125 MMF, then that's the type of fiber that should be used in the conditioned patch cable as well.

CHAPTER 11
10 Gigabit Ethernet

The IEEE 10 Gb/s standard was first specified in the 802.3ae supplement, which was adopted in 2002. This supplement defined the basic 10 Gigabit system and a set of fiber optic media standards. Subsequent 10 Gb/s supplements have added copper media types, including a short-range copper connection based on twinaxial cable and a twisted-pair media system capable of reaching 300 m. This chapter describes the copper media systems first, and then the fiber optic systems.

Table 11-1 lists several of the supplements created during the development of the 10 Gb/s Ethernet standards. These supplements have since been incorporated into the main 802.3 standard as Clause 44, "Introduction to 10 Gb/s baseband network," and Clauses 46 through 55, which describe the media systems and other elements.

Table 11-1. 10 Gigabit supplements

Supplement	Date	Identifier
802.3ae	June 2002	10 Gb/s and fiber optic media systems
802.3ak	February 2004	10GBASE-CX4 twinaxial copper media
802.3an	June 2006	10GBASE-T twisted-pair copper media
802.3aq	September 2006	10GBASE-LRM 10 Gb/s over multimode fiber

The 10 Gb/s media systems support full-duplex mode only. In the modern era of switched Ethernet ports with full-duplex media links connecting ports to devices, the normal operating mode is full duplex. The Gigabit Ethernet half-duplex mode was never adopted by vendors, and given the timing constraints, a half-duplex mode for 10 Gb/s operation made even less sense than it did for 1 Gb/s Ethernet. Therefore, it was decided at the outset that there would not be any half-duplex mode for 10 Gb/s Ethernet.

10 Gigabit Standards Architecture

The multiple specifications involved in the 10 Gb/s system are organized and defined using 802.3 physical signaling sublayers. There are four sets of physical layer (PHY) specifications, also called "families" in the standard, which are grouped by their use of the same signal encoding techniques and other elements.

Figure 11-1 provides logical diagrams of the four groups, or families, of 10 Gb/s PHY specifications, showing the sublayers. The MAC control, MAC, and reconciliation sublayers are the same across all of the groups.

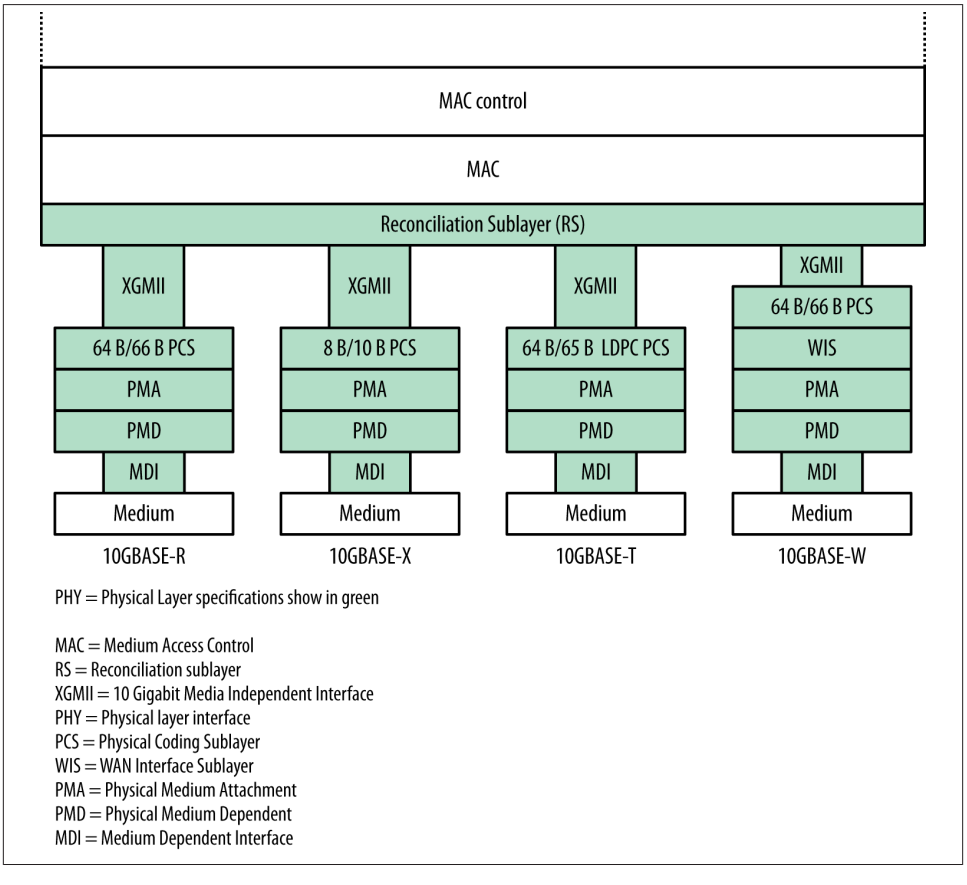


Figure 11-1. 10 Gb/s sublayer groups

The four groups are:

10GBASE-R
: Based on 64B/66B signal encoding, this includes the following optical fiber media systems: 10GBASE-SR, 10GBASE-LR, 10GBASE-ER, and 10GBASE-LRM.

10GBASE-X
: Based on 8B/10B signal encoding, including both fiber optic (10GBASE-LX4) and copper (10GBASE-CX4) media systems.

10GBASE-T
: Based on 64B/65B encoding, supporting transmission over twisted-pair cabling.

10GBASE-W
: Based on 64B/66B encoding that is encapsulated and transmitted over an OC-192 SONET optical fiber system. This includes the 10GBASE-SW, 10GBASE-LW, and 10GBASE-EW media specifications.

As sometimes happens with Ethernet standards, not all of the 10 Gigabit media systems defined in the standard have been widely adopted. The 802.3 standards provide the specifications for a range of media system technologies that have been identified as useful and likely to succeed. However, the marketplace is the proving ground for what will actually be adopted by vendors and customers.

10 Gigabit Ethernet Twisted-Pair Media Systems (10GBASE-T)

Providing 10 billion bits per second over a twisted-pair cabling system was once thought to be impossible, making the 10GBASE-T standard a major engineering achievement. The 802.3an supplement for 10GBASE-T was adopted in June 2006, four years after the initial 802.3ae 10 Gb/s standard. The 802.3an supplement has since been included in the main Ethernet standard as Clause 55, "Physical Coding Sublayer (PCS), Physical Medium Attachment (PMA) sublayer and baseband medium, type 10GBASE-T." The 10 Gigabit twisted-pair media system required a combination of signal processing and signal transmission techniques that pushed the state of the art, and that exploited the full signal-carrying capability of twisted-pair cabling.

Not surprisingly, supporting 10 Gb/s signaling over 100 meters of unshielded twisted-pair (UTP) cables required that the standard for Category 6 cabling be augmented to provide specifications to ensure adequate high-frequency performance and improved signal transmission characteristics. The new version is called Augmented Category 6 (Category 6A, or Cat6A for short). Cat6A cabling is specified in the TIA/EIA-568-B.2-ad10 standards document, and also described as Class E_A cabling in the ISO/IEC 11801 standard. Twisted-pair cabling, including Cat6A, is described in more detail in Chapter 16.

10GBASE-T Signaling Components

A 10GBASE-T computer interface or switch port includes a built-in transceiver (PHY) and the medium dependent interface (MDI) used to make a direct connection to the 10GBASE-T twisted-pair segment. Switch ports have built-in Ethernet interfaces. A computer may have an Ethernet interface built in at the factory, or the interface may be provided on an adapter card that is installed into the computer system. There is no transceiver interface exposed to users for connections to external transceivers in the 10GBASE-T system.

Figure 11-2 shows a desktop computer connected to a switch port with a Cat6A cable that is capable of operating at 10 Gb/s. Network interfaces in computers and switch ports usually provide a combination of transceiver electronics to support operation at multiple speeds. Three speeds are usually supported, and 10 Gb/s twisted-pair interfaces typically operate at 100 Mb/s, 1000 Mb/s, and 10 Gb/s speeds. The Clause 28 Auto-Negotiation standard for twisted-pair media is used to automatically configure the speed of operation over the link.

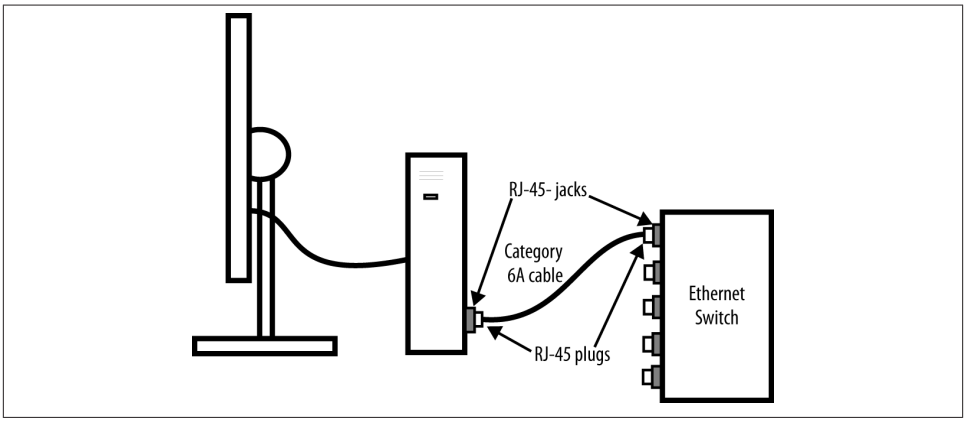


Figure 11-2. 10GBASE-T Ethernet interface

A 10GBASE-T connection uses all four pairs for signaling. The automatic MDI/MDI-X system for signal crossover, originally developed in the 1000BASE-T system, is used to achieve correct signal crossover between interfaces connected over a cable segment. The 10GBASE-T PHY specifications include a training phase during which signals are sent that allow the transceivers to automatically detect and correct for swapped wire pairs in the cabling and swapped signal polarity within a given wire pair.

10GBASE-T Signal Encoding

The 10GBASE-T system requires a complex set of *digital signal processing* (DSP) techniques to provide 10 gigabits per second across twisted-pair cabling plants with cables that support a bandwidth of 250 MHz (Cat6) or 500 MHz (Cat6A). The maximum distance of 100 meters is achievable over UTP Cat6A cabling, while the less-capable Cat6 cabling can be expected to support shorter distances up to 55 m, depending on the quality of the cable installation. To provide 10 Gb/s transport of Ethernet frame data across the link, each of the four signal pairs effectively delivers 2.5 Gb/s in both directions simultaneously.

The 10GBASE-T physical coding sublayer (PCS) defines a signaling system that couples a 10 Gigabit media independent interface (XGMII) to the physical medium attachment (PMA) sublayer. The XGMII provides a 32-bit-wide parallel signal bus to carry Ethernet frame bits, providing 32 bits of frame data at a time. The 10GBASE-T PCS works on eight octets at a time received from the XGMII (two transfers of 32-bit-wide data), and these 64-bit chunks of data are encoded as 65-bit data blocks (64B/65B).

The blocks of data are transformed into a line code that reduces the required bandwidth over the cabling, using a baseband signaling system called 16-level pulse amplitude modulation (PAM), which reduces the signaling transitions to 800 Mbaud. The bits are further transformed using another bandwidth-reducing line code called 128-DSQ (double square 128), which limits the bandwidth needed to transmit 10GBASE-T signals over the cabling to less than 500 MHz.

A powerful forward error correction system using a code called low-density parity check (LDPC) helps to minimize signaling errors, and ensures nearly error-free operation at close to the fundamental limits of information-carrying capability for this type of cabling. The worst-case bit error rate for this system is specified as 10 to the minus 12 (10^{-12}), which means no more than 1 bit error in every trillion bits sent, on average. Most twisted-pair systems operate at a significantly better error rate, and it is common to see Ethernets run for long periods of time with no errors reported.

This section is intended to provide a brief overview of the signaling system. Anyone who needs to know all of the details of the 10GBASE-T signal encoding system will find them described in depth in Clause 55 of the 802.3 standard.[1]

Signaling and data rate

To transmit and receive data on all four wire pairs simultaneously, the 10GBASE-T transceivers at each end of the link contain four identical transmit sections and four

1. A description of the encoding components of the PHY can also be found in a white paper from the Ethernet Alliance entitled "10 Gigabit Ethernet on Unshielded Twisted-Pair Cabling."

identical receive sections. Each of the four wire pairs in the link segment is connected to both transmit and receive circuitry in the transceiver.

Figure 11-3 is a schematic drawing of two transceivers connected together over four twisted pairs of wire. All four wire pairs simultaneously send and receive data, and 3.5 bits of Ethernet data are encoded and sent per signal transition on each wire pair.

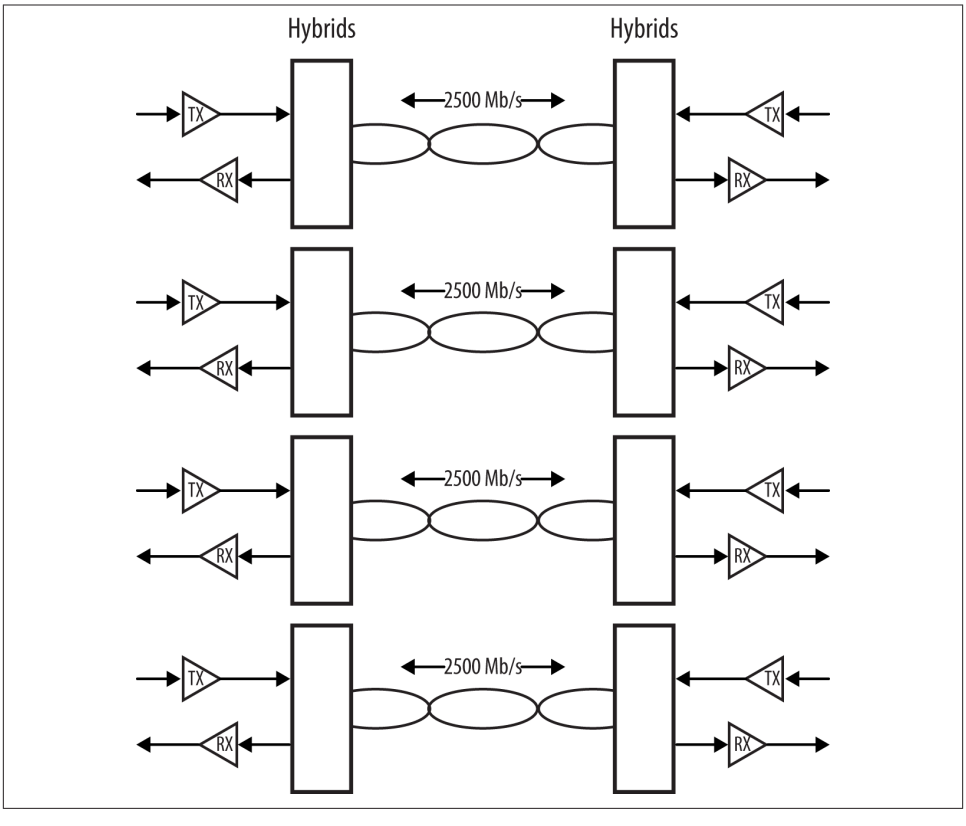


Figure 11-3. 10GBASE-T signal transmission

The continuous signaling in both directions on all four wire pairs generates signal echo and crosstalk, which the 10GBASE-T system handles with a further set of digital signal processing techniques. These include echo cancellation, *near-end crosstalk* (NEXT) cancellation, and *far-end crosstalk* (FEXT) cancellation. Another DSP technique is *signal equalization*, to help compensate for signal distortion over the channel. The transceiver also includes a self-synchronizing scrambler to spread out the electromagnetic emission patterns in the data to help avoid signal emissions from the cable.

Figure 11-3 shows the signals passing through an electronic circuit element called a *hybrid*, which provides simultaneous bi-directional transmission with echo cancella-

tion. The output voltage of the transmitter into the 100 ohm impedance of the twisted-pair cable is approximately 2 volts, peak to peak.

Signal clocking

Auto-Negotiation is used to support the configuration of signal clocking, and is embedded in the transceiver. To help improve signal processing, 10GBASE-T includes a *master-slave system* of synchronous signal clocking for each wire pair. The master and slave ends of a given wire pair synchronize their signaling using the clock signal provided by the master. This allows the transceiver circuits linked over each wire pair to differentiate between the signals they are sending and signal echo, crosstalk, and "alien" signals sent on other pairs. This, in turn, makes it possible to strongly suppress these interfering signals, thus improving the signal-to-noise ratio. The Auto-Negotiation mechanism is used to decide which transceiver becomes the master.

The signal encoding scheme provides both data symbols and symbols used for control and other purposes. These other symbols include the IDLE symbol, which is continually sent when no other data is present. Energy Efficient Ethernet is supported and can be used to save power during periods when there is no data to send, as described in Chapter 7.

10GBASE-T cabling requirements

The high rate of signaling and the complex signaling techniques result in an increased sensitivity to signal impairment issues on twisted-pair cable segments. Older cable types cannot provide the signal quality needed for 10GBASE-T signals, and for that reason, the IEEE standard does not specify operation over Cat5e cables. For best results, use cables with a high signal quality rating, such as Category 6A.

Reliable 10 Gigabit Ethernet operation requires that all patch cords be correctly assembled using high-quality components. The twisted pairs must maintain their twists as close as possible to the RJ45 connectors, and the connectors must be high quality for the best signal-carrying capabilities. To achieve this level of signal quality, you must purchase high-quality patch cords, manufactured under carefully controlled conditions, that are tested and rated to meet Category 6A specifications.

10GBASE-T Media Components

Although shielded cabling is also used in cabling systems, especially in European countries, the 10GBASE-T standard has a goal of providing operation over unshielded twisted-pair cabling. The vast majority of the cabling systems in the United States use unshielded cable.

The following cabling types can be used to build a 10GBASE-T twisted-pair segment:

- Class F, Catgeory 7 shielded cable, as specified in ISO/IEC 11801. This shielded cabling exceeds the minimum requirements for 10GBASE-T segments up to 100 m (328.08 feet).
- Class E_A, Augmented Category 6 (Category 6A), both shielded and unshielded, as specified in the ISO/IEC 11801 Edition 2.1 and TIA-568-C.2 standards. These cabling types exceed the minimum requirements for 10GBASE-T segments up to 100 m.
- Class E, Category 6, screened (shielded) cable, as specified in ISO/IEC TR-24750 and TIA/EIA TSB-155. These shielded cabling types exceed the minimum requirements for 10GBASE-T segments up to 100 m.
- Class E, Category 6, unscreened (unshielded), as specified in ISO/IEC TR-24750 and TIA/EIA TSB-155. There is a 55 m (180.44 foot) limit on length, but there is no guarantee that you can achieve 55 m of error-free operation on Cat6 UTP cabling; it depends on the quality of the cable and the cabling installation.

  According to TSB-155, Category 6 UTP cabling should be able to provide maximum channel lengths of between 37 and 55 meters, depending on the amount of alien crosstalk found in the cabling system. Testing the cable segment to determine the signal quality is recommended to ensure correct operation. There are a number of mitigation techniques listed in the TIA document TSB-155 that can be used to improve signal quality on a given segment, including minimizing the number of cable interconnects in any given cable segment and replacing Cat6-rated connectors with Category 6A connectors.

Category 5e is not supported. The 10GBASE-T PHY includes a signal training phase that operates between transceivers over a link, which provides information about the signal quality of the link cable segment. This information allows the transceivers to detect cable signal quality that does not meet the requirements for reliable 10GBASE-T operation. In response, transceivers can use Auto-Negotiation to negotiate a lower-speed mode of operation that can be supported by the cable.

The 10GBASE-T transceiver is required to do a significant amount of signal processing to provide 10 Gb/s signaling, which requires a lot of integrated circuit components. Because of the number of circuits required, the first generation of 10GBASE-T transceivers consumed roughly 10 watts of power.

Each new generation of 10GBASE-T transceiver has been smaller and more efficient, as the circuits have been improved and reduced in size. Newer transceivers based on 40 nm chip technology consume from 2.5 to 4 watts , which is less power than the previous generation. The fourth generation of transceivers, based on 28 nm technology, was announced as this was being written in late 2013, and they consume 1.5 watts when

supporting a 100 m segment over Cat6A cabling.[2] A shorter cable distance can reduce the power level even more, as described in "10GBASE-T Short-Reach Mode" on page 181. Operating the link with Energy Efficient Ethernet mode enabled can reduce the power consumption even further.

Given that the SFP+ module can only support components with a maximum power consumption of up to 1.5 watts, it has not been possible to build 10GBASE-T SFP+ transceiver modules based on previous generations of transceiver technology. Instead, 10GBASE-T switch ports or interface adapter cards are provided as "fixed ports"; there are no pluggable 10GBASE-T modules for SFP+ switch ports. The newest 10GBASE-T PHYs make it possible to provide 10GBASE-T ports that consume less power, and may also make it possible for vendors to provide SFP+ 10GBASE-T modules for switch ports.

Eight-position RJ45-style modular connectors that meet or exceed Category 6A specifications are recommended for use on a 10GBASE-T twisted-pair segment, for operation to 100 m.

Eight-position RJ45-style jack connectors

The 10GBASE-T media system uses four pairs of wires that are terminated in an eight-position (RJ45-style) connector. A 10GBASE-T system uses four pairs of wires, so all eight pins of the connector will be used.

As shown in Table 11-2, the four wire pairs are used to carry four bi-directional data signals (BI_D). The four bidirectional data signals are called BI_DA, BI_DB, BI_DC, and BI_DD. The data signals on each pair of a 10GBASE-T twisted-pair segment are polarized, with one wire of each signal pair carrying the positive (+) signal, and the other carrying the negative (–) signal. The signals are connected so that both wires associated with a given signal are members of a single wire pair.

Table 11-2. 10GBASE-T RJ45 signals

Pin number	Signal
1	---- BI_DA+ ----
2	---- BI_DA- ----
3	---- BI_DB+ ----
4	---- BI_DC+ ----
5	---- BI_DC- ----
6	---- BI_DB- ----
7	---- BI_DD+ ----
8	---- BI_DD- ----

2. Press release on the new 28 nm 10GBASE-T PHY.

The 10GBASE-T transceivers include circuits that can detect incorrect signal polarity (polarity reversal) in a wire pair. These circuits can correct polarity reversal by automatically moving the signals to the correct circuits inside the transceiver. However, it is not a good idea to depend on this ability. Instead, all cables should be wired so that correct signal polarity is observed.

10GBASE-T Link Integrity Test

The 10 Gigabit Ethernet transceiver circuits continually monitor the receive data path for activity as a means of verifying whether the link is working correctly. The signaling system used for 10GBASE-T segments continually sends signals—even during idle periods when there isn't any traffic on the network. Therefore, activity on the receive data path is sufficient to provide a check of link integrity.

10GBASE-T Configuration Guidelines

The Ethernet standard contains guidelines for building a single 10GBASE-T twisted-pair segment, shown here in Table 11-3. Annex 55B of the standard provides extra guidelines for mitigating signal crosstalk, including recommendations to loosen cable bundle bindings and to minimize the distance that cables run in parallel. Annex 55B notes that "The star wiring topology, where the cables are distributed radially from a centralized telecommunications closet to each work area, reduces the distances over which link segments are in close proximity."[3]

Table 11-3. 10GBASE-T single segment guidelines

Media type	Maximum segment length	Maximum number of transceivers (per segment)
Cat6A unshielded twisted-pair 10GBASE-T	100 m (328.08 feet)	2
Cat6 unshielded twisted-pair 10GBASE-T	Up to 55 m (180.4 feet) possible	2
Cat5e unshielded twisted-pair 10GBASE-T	Not supported	

The ISO/IEC 11801 specifications include minimum length specifications for components of an unshielded twisted-pair segment, to minimize signal crosstalk issues. The specifications provide a two-meter minimum length for a patch cord connecting to an Ethernet interface to minimize signal crosstalk issues. If shielded or "screened" twisted-pair cabling is used, then the crosstalk issues do not arise due to the higher-quality signal performance of shielded cable.

The 10GBASE-T specification allows a segment with a maximum length of 100 meters. A 10GBASE-T segment is unlikely to function well over cable distances longer than 100 meters due to the increased signal impairments that occur over longer cable lengths.

3. IEEE Std 802.3-2012, Annex 55B, p. 732.

10GBASE-T Short-Reach Mode

Given the complex signal processing needed to transmit 10 Gb/s signals over twisted-pair cabling, the power consumption of the electronics required to provide the signal processing is a concern on 10GBASE-T ports. To help reduce the power requirements, an optional short-reach mode was defined to provide a lower-power means for operation on a high-quality cable segment consisting of Category 6A or Class F cable up to 30 m in length. Vendors have reported that their 10GBASE-T PHYs can consume up to 60% less power when running in short-reach mode.

Short-reach mode is automatically detected between compliant transceivers, and short-reach capability is signaled with Auto-Negotiation. When the channel length is 30 m (98.4 feet) or less, the 10GBASE-T transceivers are able to reduce their power dissipation while still maintaining the specified bit error rate performance.

Transmit power is reduced in short-reach mode, and some of the echo cancellation and signal filtering can also be powered down internally in the transceiver. Studies have found that across a number of facilities studied, 37% of the cable segments are less than 30 meters in length.[4] The number of short segments will vary from facility to facility, but in general there should be enough short segments in many data centers and other facilities to make it possible to save power by using short-reach mode.

10GBASE-T Signal Latency

The amount of digital signal processing required for a 10GBASE-T transceiver to function necessarily imposes some signal delay as the signal travels through the various components, such as filters and equalizers. This delay, combined with the unavoidable serialization delay over 100 m of cable, can result in as much as 2.5 microseconds (2.5 µs) of latency for a 10GBASE-T segment. In short-reach mode, the latency can be reduced to about 1.5 microseconds by reducing the number of circuits required to process the signal, and by reducing the length of the cable.

This can be compared to the latencies of approximately 1.5 microseconds seen on short fiber optic links or short copper connections such as the direct attach cables that are described later in this chapter.

The majority of computers or servers will not be affected by the few extra millionths of a second of latency that occur on a long 10GBASE-T segment. Normal communication delays between computers in a data center are affected by many factors, such as processor workload, memory access, and disk access. Software operations typically contribute latency on the order of milliseconds (ms). As each millisecond is 1,000 microseconds,

4. See Valerie Maguire and David Hess's BICSI presentation, "The 40Gbps Twisted-Pair Ethernet Ecosystem."

these normal delays swamp the few extra microseconds involved in 10GBASE-T operations.

However, some specific scenarios, such as a high-performance computing cluster or a central file server or database server that supports a number of other servers in a data center, are extremely sensitive to any extra latency. In these cases, you may wish to avoid any extra latency by connecting these application servers to 10 Gb/s ports with direct attach cables or fiber optic segments.

10 Gigabit Ethernet Short Copper Cable Media Systems (10GBASE-CX4)

The 10GBASE-CX4 short-reach copper cable media system was initially defined in the 802.3ak supplement to the standard, which was adopted in 2004. The CX4 specifications, adopted into the standard as Clause 54, define a media system based on twinaxial cables of a type that is also used in the Infiniband high-speed network technology. The twinaxial cable is similar to coaxial cable, but with two inner conductors instead of one. Twinaxial is capable of carrying high-speed signals for short distances of up to 15 m.

The 10GBASE-CX4 standard defines a medium dependent interface that uses a fairly large 16-pin connector, adopted from the Infiniband standard, measuring approximately one inch wide and four-tenths of an inch thick. A media segment that is compliant with the 10GBASE-CX4 standard must use this connector type on the cable. The cable and its connectors are sold as an assembly of a fixed length, usually in the range of 1 m to 15 m.

Figure 11-4 shows the 10GBASE-CX4 cable assembly, which comes with the 16-pin connectors permanently attached to each end. The 10GBASE-CX4 standard, with its higher-performance cable and the reduced electronics required to send a signal, provides the advantage of lower power than a 10GBASE-T transceiver. This also provides lower cost and lower signal latency through the transceiver. However, disadvantages include a large connector and a relatively stiff and bulky cable with multiple twinaxial conductors to carry the four lanes of signals.

Figure 11-4. 10GBASE-CX4 cable assembly

The 10GBASE-CX4 connector and cable size was not popular with vendors, and this standard has not been widely adopted by the marketplace. Consequently, there are very few 10GBASE-CX4 Ethernet products available. Given this limited availability, we will not describe the 10GBASE-CX4 system in detail.

Instead, we will describe the twinaxial cable variant that was developed by vendors using a different set of signaling specifications and replacing the large CX4 connector with a smaller SFP+ connection module.

10 Gigabit Ethernet Short Copper Direct Attach Cable Media Systems (10GSFP+Cu)

The 10GSFP+Cu media type is not specified in the 802.3 standard, and the shorthand identifier was invented by the vendors who developed this cable type. This low-cost short copper cable segment is useful, for example, for interconnecting a stack of switches, and for short-distance connections between switch ports and Ethernet interfaces on servers and other devices. Prior to the development of 10GBASE-T, this direct attach cable (also referred to as a DA or DAC) was the only low-cost copper connection available for 10 Gb/s operation.

The direct attach cable is terminated with a small form-factor (SFF) connector module called the SFP+ (small form-factor pluggable plus). The SFP+ transceiver module fits into a port that is roughly the size of an RJ45-style port, making it possible for vendors to provide a higher port density on their switches.

The SFP+ module is not standardized by any official standards body, but instead is specified by a multisource agreement (MSA) that was developed among competing

manufacturers.[5] Multisource agreements are the primary method used to develop communications connectors and transceiver modules that are used in Ethernet and other network systems. As technology advances, cabling and equipment vendors work together to develop smaller and more efficient connectors and modules, using the MSA as a rapid method of developing these enhancements. The SFP+ is the second generation of the SFP standard. The original SFP specifications support operation up to 4.5 Gb/s. The SFP+ features support for 10 Gb/s or higher signaling, based on improved impedance matching specifications for better signal transmission.

Vendors provide both active and passive versions of the 10GSFP+Cu DA cable. A direct attach cable assembly is considered active if there are signal processing electronics in the SFP+ module to improve signal quality and provide a longer cable distance. The least expensive approach is the passive direct attach cable, which tends to be shorter; the active direct attach cable makes it possible to support longer and thinner cable assemblies. The available cable lengths range from 1 m to 7 m, although you will find variations in supported cable lengths among different vendors.

The direct attach twinaxial cables use the same SFP+ connector module that is used on 10 Gb/s optical fiber links. However, rather than using an optical transceiver at each end and a length of fiber optic cable, the direct attach cable uses the SFP+ module, leaving out the expensive optical lasers and other electronic components. In both active and passive cables, a small electrical component is used to identify the SFP+ module and cable type to the Ethernet interface; that component is low cost and consumes very little power.

A direct attach cable is a fixed assembly that is purchased at a given length, with the SFP+ modules permanently attached to each end of the cable. Although significantly thinner than 10GBASE-CX4 cable, this is still a relatively stiff cable. To install the cable on anything other than closely associated equipment, you must be able to route the cable and its attached SFP+ modules through any intervening cable management trays and cable guides.

10GSFP+Cu Signaling Components

An SFP+ port may support either active or passive DA cables, or both. There is no standard for this cable type, so you cannot assume that a direct attach port can support either cable type. Instead, you must refer to the documentation for the interface or switch port to determine what kind of cables the port can support.

When an SFP+ port supports a direct attach connection, all you need to do is insert the SFP+ module on the end of the DA cable into the port until it latches. The SFP+ active

5. The specifications for SFP+ twinaxial transceivers include SFF-8431 (passive cable) and SFF-8461 (active cable) and can be found on the SFF Committee website.

and passive cable assemblies are hot pluggable, meaning that you can safely connect and disconnect them with power enabled on the switch or computer interface.

Figure 11-5 shows a set of four SFP+ ports on a switch, an SFP+ direct attach cable, and an Ethernet interface with two SFP+ DA ports. The twinaxial cable has permanently attached SFP+ modules on each end, and the SFP+ modules are plugged into the SFP+ ports on the switch and the Ethernet interface. In this figure, there are also RJ45 ports on the left side of the switch, which provides an opportunity to see how the RJ45 and SFP+ ports compare in size.

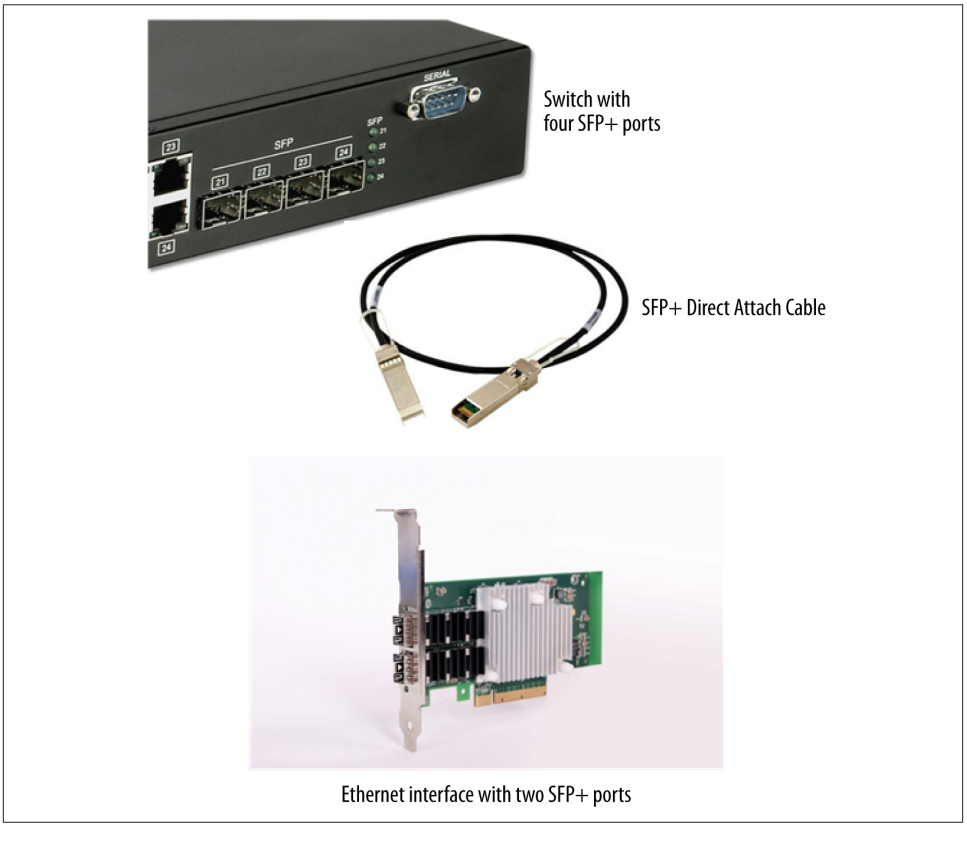


Figure 11-5. 10GSFP+Cu connection components

Given the lack of a formal 802.3 standard for this cable type, vendor and cable interoperability is not guaranteed.[6] To ensure interoperability, one approach is to purchase cable assemblies from the same vendor that provides the switch port. Another approach is to request a list of verified cable assemblies from the switch and interface vendors, to ensure that the direct attach cables that you buy are the correct type (active or passive), and have been tested and verified to work with the Ethernet interfaces you are connecting together.

10GSFP+Cu Signal Encoding

The direct attach cables and their SFP+ connector modules use an electrical signaling interface called *SFI*, which is defined as the "SFP+ high-speed serial electrical interface." The SFI definition enables 10 Gb/s operation over a single differential signaling path in each direction for a total of two pairs, or four wire connections. The twinaxial cable includes the two pairs of signal-carrying wires in a coaxial cable format, providing high performance and stable signaling over the length of the cable.

When in SFI operating mode, an Ethernet interface uses the 10GBASE-R physical coding sublayer specifications and the 10 Gigabit physical medium attachment specifications that are defined in 802.3 Clauses 49 and 51, respectively. The SFI specification provides a full-duplex electrical interface that uses a single self-clocked serial differential link in each direction to achieve 10 Gb/s data throughput.

The serial link transfers scrambled data at 10.3125 Gbaud to accommodate both data and the overhead associated with 64B/66B coding. The self-clocked nature eliminates skewing between clock and data signals. The SFI link requires a 100 ohm cable impedance, while the signal termination electronics provide both differential and common-mode suppression of signal noise and reflections. The direct attach copper cable is designed to meet a bit error rate specification of 10^{-12}, which is a potential for 1 bit error in every trillion bits sent.

The SFP+ MSA specifications for direct attach cables note that 10GSFP+Cu connections can only be used on systems with common power grounds. The power supplies for the Ethernet switches, and for any computers or other switches that are connected to them over direct attach cables, must themselves be connected to the same local power grid with a common ground between all devices, as a difference in ground potentials between systems connected with DA cables could result in electrical damage to the interfaces or the devices.

6. In 2009, the Ethernet Alliance held an "interoperability plugfest" for direct attach cables to demonstrate multi-vendor interoperability for this cable type. The report on the results of the plugfest is entitled "SFP+ Direct Attach Copper Interoperability Demonstration White Paper."

10GSFP+Cu Link Integrity Test

The 10 Gigabit Ethernet transceiver circuits continually monitor the receive data path for activity as a means of verifying whether the link is working correctly. The signaling system used for 10GSFP+Cu cable segments continually sends signals—even during idle periods where there isn't any traffic on the network. Therefore, activity on the receive data path is sufficient to provide a check of link integrity.

10GSFP+Cu Configuration Guidelines

The direct attach cables used in the 10GSFP+Cu system come in a limited range of sizes, with lengths in the range from 0.6 m (1.96 feet) to 7 m (22.96 feet) being the most widely sold. It is up to you to verify which cable lengths are supported by your vendor, and whether the cable can be active or passive. Some vendors support both types of cables on their SFP+ ports, and others do not.

The twinaxial cable itself is relatively stiff, and may present difficulties if you need to route it through tight spaces. There are two conductors in each twinaxial cable, with two twinaxial cables inside a common sheath providing a total of four conductors for signals. As an example, one vendor's direct attach cable has an outer diameter of 0.180 inches, or roughly 3/16 inch. The minimum bend radius for this cable is one inch. If you bend the cable any tighter than this, it may affect the signal quality.

10 Gigabit Ethernet Fiber Optic Media Systems

The 802.3ae supplement that initially defined 10 Gigabit Ethernet was adopted in 2002; it defines a set of fiber optic media standards based on the physical layer signaling specifications shown in Figure 11-1.

There are a number of 10 Gigabit fiber optic media system specifications, which provide short-range operation over multimode fiber optic (MMF) cable and long-reach operation over single-mode fiber optic (SMF) cable. The physical layer specifications for these systems are grouped as local area network (LAN) PHYs. The standard groups the optical specifications as 10GBASE-S for the short-reach systems, and 10GBASE-L for the long-reach systems.

There are also 10 Gigabit fiber optic media systems designed to be carried over wide area network (WAN) fiber optic systems based on the synchronous optical network (SONET) standard. These are grouped as WAN PHYs.

A further complication is that a number of pluggable fiber optic transceivers have also been developed, beginning with the earliest XENPAK transceivers, followed by the closely related X2 module, and proceeding to the XFP and the currently popular SFP+ modules. These connector types are all based on multisource agreement specifications developed by vendors.

When you purchase 10 Gigabit fiber optic transceivers, you need to make sure that you are buying the correct type to fit the switch port or computer interface connection involved. To do that, you need to consult the vendor documentation.

Figure 11-6 shows several types of optical transceiver modules that have been used for 10 Gb/s optical media types. The optical fiber plugs into the modules, using either an SC optical connector (XENPAK and X2) or LC optical connectors (SFP+).

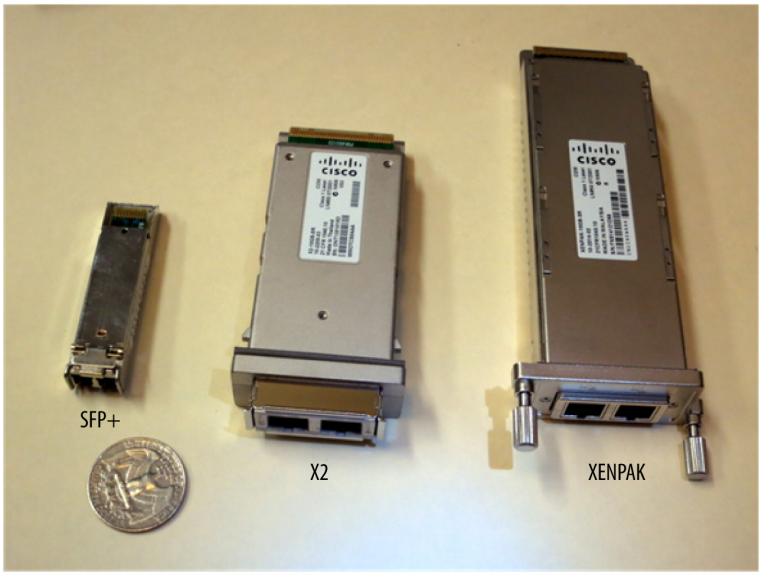
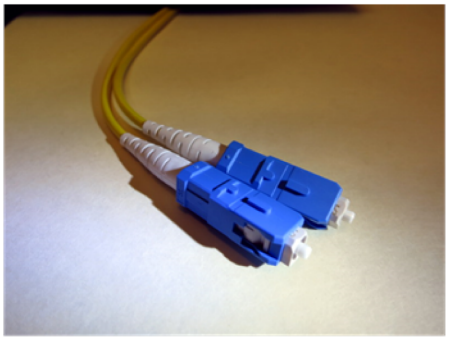


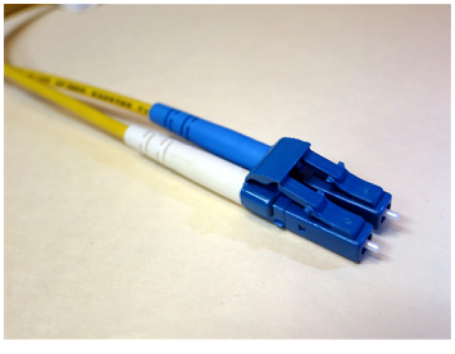
Figure 11-6. 10 Gigabit optical transceiver modules

Optical cables can be ordered in specific lengths and with specific connector types, as required by your circumstances. If you are connecting an optical jumper cable from a switch port directly to a server interface, then you need a short cable with the appropriate cable connector on each end. If you are connecting a switch port to an optical connector termination point in your data center or cabling closet, the optical connectors in your termination point might be different than the ones used on the 10 Gb/s transceiver module, which would require a cable with a different type of optical connector on each end. Fiber optic cabling and components are described in more detail in Chapter 16.

Figure 11-7 shows the SC and LC optical connectors attached to fiber optic cables.



SC connector used with XENPAK and X2



LC connector used with XFP and SFP+

Figure 11-7. 10 Gigabit optical cable connectors

10 Gigabit LAN PHYs

The 10 Gigabit LAN PHYs include five media systems, detailed in Table 11-4.

Table 11-4. 10 Gigabit LAN PHYs and signal encoding

Media type	Physical coding sublayer	Optics	Comment
10GBASE-SR	64B/66B 10GBASE-R	10GBASE-S	Short reach, serial over MMF
10GBASE-LX4	8B/10B 10GBASE-X	10GBASE-S and 10GBASE-L	Short reach over MMF, long rlong reach over SMF
10GBASE-LR	64B/66B 10GBASE-R	10GBASE-L	long reach over SMF
10GBASE-LRM	64B/66B 10GBASE-R	10GBASE-L	Long reach over MMF
10GBASE-ER	64B/66B 10GBASE-R	10GBASE-L	Extended reach over SMF

Each of the LAN PHYs was specified to provide a media system for a specific set of uses. The short-reach systems are intended for connections between switches in a building or a data center, and from switch ports to server interfaces. The long-reach systems, intended for backbone links on campus and enterprise networks, use more expensive

single-mode laser-based optics to drive 10 Gigabit signals over longer distances, with ranges from 10 km for the LR system up to 30–40 km for the ER system.

Let's take a closer look at each of these media systems:

10GBASE-SR
: Designed for short-reach applications, this media type operates over a single pair of multimode fiber optic cables that meet the 10GBASE-S fiber optic specifications. Transceivers for this media type have used the X2 module, but the most popular current form is the SFP+ module.

  There is a nonstandard variant of this media type called 10GBASE-SRL, for "short reach lite." This variant is available from a few vendors, and has a limited segment length of 100 m. The goal is to provide a low-cost connection using less expensive short-reach optics.

10GBASE-LX4
: This media type is designed to operate over both single-mode and multimode fiber optic cables using four separate laser light sources. The LX4 system has its own set of fiber optic specifications. In the 10GBASE-LX4 media system, the Ethernet frame data is split into four lanes, each operating at 3.125 Gb/s using the 8B/10B encoding scheme and other signaling elements defined in the 10GBASE-X physical coding sublayers. The four lanes of frame data are transmitted over the cable using a system called *coarse wave division multiplexing* (CWDM). Each of the lasers uses a unique wavelength of light around 1,310 nm, with all four wavelengths simultaneously transmitted over a single pair of fiber optic cables. This media type is relatively complex and expensive, and was not widely adopted in the marketplace.

10GBASE-LR
: Designed for long-reach applications, this media type operates over a single pair of single-mode fiber optic cables that meet the 10GBASE-L specifications, using laser light sources that transmit at the 1,310 nm wavelength. Signal encoding is based on the 10GBASE-R PCS, uses the 64B/66B block encoding scheme, and transmits data in a single serial stream at 10.3125 Gb/s.

10GBASE-LRM
: The long-reach multimode (LRM) media type is designed to operate over multimode fiber optic cables that meet the 10GBASE-S fiber optic specifications, using 1,310 nm laser light sources. It transmits at a line rate of 10.3125 Gb/s using the 10GBASE-R PCS and 64B/66B block encoding. This media type can support segment lengths of up to 220 m over the older FDDI-grade multimode fiber, and the same distance over the OM1, OM2, and OM3 multimode fiber types.

  A mode-conditioning patch cable is required for connection to FDDI grade and OM1 or OM2 cables to ensure that the maximum segment distance can be achieved,

as discussed at the end of Chapter 10. No mode-conditioning patch cable is required for OM3 and OM4 cable types.

The letters "OM" stand for *optical multimode*. The OM specifications are defined in international standard ISO/IEC 11801.

10GBASE-ER
: The extended reach media type operates over a single pair of single-mode fiber optic cables, using 1,510 nm laser light sources. It transmits at a line rate of 10.3125 Gb/s using the 10GBASE-R PCS and 64B/66B block encoding.

10 Gb/s Fiber Optic Media Specifications

The fiber optic media segments require two strands of cable: one for transmitting and one for receiving data. The required signal crossover, in which the transmit signal (TX) at one end is connected to the receive signal (RX) at the other end, is performed in the fiber optic link.

Maximum segment lengths are dependent on a number of factors. Fiber optic segment lengths will vary depending on the cable type and wavelength used by the media type. More information on multimode and single-mode fiber optic segments and components can be found in Chapter 17.

The 10 Gigabit standard provides specifications that the multimode and single-mode fiber optic media must meet to support 10 Gigabit Ethernet over various lengths of cable, listed in Table 11-5. The standard groups the cable specifications together as 10GBASE-S for short-reach multimode and 10GBASE-S for long-reach single mode.

Table 11-5. Optical specifications for 10GBASE-S

Fiber type	Minimum modal bandwidth at 850 nm (MHz-km)	Channel insertion loss (dB)	Operating range (m)
62.5 µm MMF	160	1.6	2 to 26
62.5 µm MMF (OM1)	200	1.6	2 to 33
50 µm MMF	400	1.7	2 to 66
50 µm MMF (OM2)	500	1.8	2 to 82
50 µm MMF (OM3)	2,000	2.6	2 to 300
50 µm MMF (OM4)	4,700	2.9	2 to 400

Multimode fiber optics are less expensive than single-mode optics and are able to transmit light for relatively short distances. The *modal bandwidth* of a multimode fiber optic

cable refers to its signal-carrying characteristics. Higher modal bandwidth means that the cable can transmit a signal for a longer distance while maintaining sufficient signal quality to the optical receiver in the Ethernet interface at the end of the link. The 62.5 µm and 50 µm fiber types refer to the diameter of the signal-carrying portion of the fiber optic cable.

The total link segment power budget for each 10GBASE-S fiber type is 7.3 dB, with varying amounts of power consumed by optical noise characteristics, intersymbol interference issues, and the like, depending on the type of the multimode fiber. The channel insertion loss is that portion of the link segment optical power budget that can be consumed by the optical cabling and connectors used on a given segment. As long as the optical power loss, as measured through the link segment from end to end, is at or below the level shown for the channel insertion loss, then the segment will operate correctly.

The 10GBASE-LX4 media system was designed to operate over both multimode and single-mode fiber optic cables, with the specifications shown in Table 11-6.

Table 11-6. Optical specifications for 10GBASE-LX4

Optical type	Modal bandwidth (MHz-km)	Channel insertion loss (dB)	Operating distance
62.5 µm MMF	500	2.0	300 m
50 µm MMF	400	1.9	240 m
50 µm MMF	500	2.0	300 m
10 µm SMF	n/a	6.2	10 km

The specifications for long-reach and extra-long-reach optical fibers used in the 10GBASE-LR and 10GBASE-ER media systems are simpler, because single-mode fiber optic cables' transmission characteristics are different than multimode, and do not involve modal bandwidth considerations.

The specifications (listed in Table 11-7) are designed to be conservative and cover the worst case of fiber optic operation. However, because single-mode fiber optic cables are designed to transmit signals for long distances, they may be able to exceed the distances shown here.

Table 11-7. Optical specifications for 10GBASE-L and 10GBASE-E

Optical type	Wavelength (nm)	Channel insertion loss (dB)	Minimum range
10GBASE-L	1,310	6.2	2 m to 10 km
10GBASE-E	1,550	10.9	2 m to 30 km[a]

[a] Operation up to 40 km is possible for "engineered" links, in which the signal dispersion characteristics of the cable meet the specifications listed in Table 52-24 of the 802.3 standard.

10 Gigabit WAN PHYs

The 10 Gigabit standards include a set of wide area network PHYs that are designed to couple a 10 Gb/s Ethernet interface to a WAN interface using a specific type of synchronous optical network technology called *SONET STS-192c*. SONET technology has been widely used to build the long-distance connections that Internet service providers use in their network systems. By developing a connection between standard Ethernet and a SONET WAN interface, the standard made it possible to build wide area network links that could connect directly to 10 Gb/s Ethernet interfaces.

The WAN PHY is based on a WAN interface sublayer (WIS). The WIS allows the 10 Gb/s WAN PHY to generate Ethernet data streams that are mapped directly to STS-192c or VC-4-64c streams at the PHY level. In the WAN PHY, the PCS data stream (including IDLE symbols) is combined with SONET path overhead in the correct order and is mapped into a standard STS-192c payload envelope. The STS-192c payload envelope is combined with the SONET line and section overhead and is mapped into the WIS frame. This is not intended to be a fully featured SONET/SDH interface. Instead, it operates as a "lightweight" version of OC-192 SONET running at the rate of 9.58464 Gb/s.

The slight speed mismatch may result in port buffer congestion in an Ethernet switch, as a stream of Ethernet frames arriving in the switch at 10 Gb/s is transmitted over the WAN interface at the slightly slower rate. In that case, the switch can use PAUSE frames in an attempt to throttle Ethernet frame transmission from sending devices. Depending on the offered frame rate and switch port buffer sizes, this could help avoid dropped frames due to congestion on a WIS interface connection to a SONET link.

The WAN PHY uses the same 10GBASE-S, 10GBASE-L, and 10GBASE-E optical specifications as the LAN PHYs, and the media systems are designated as 10GBASE-SW, 10GBASE-LW, and 10GBASE-EW. A maximum distance of up to 80 km can be supported, depending on the fiber employed. Both XENPAK and XFP modules have been developed to provide WAN PHY transceivers. There are also "multirate" modules that can support, for example, both the 10GBASE-LR and 10GBASE-LW modes of operation.

Most enterprise and campus networks will use LAN PHY versions of 10 Gigabit Ethernet, but if there is a need for a direct connection between an Ethernet switch or router interface over a link composed of SONET technology, the WAN PHYs can provide a solution.

CHAPTER 12
40 Gigabit Ethernet

The development of a new Ethernet system that operates faster than 10 Gb/s began with a Higher Speed Study Group "Call For Interest" meeting in July 2006. In response, an IEEE task force was formed to develop a 100 Gb/s Ethernet system in the 802.3ba supplement. This effort was then expanded to include 40 Gb/s Ethernet, and the combined 40 and 100 Gb/s 802.3ba supplement was completed and published in 2010. The 40 and 100 Gb/s specifications were adopted into the standard as Clauses 80 through 89.

It's unusual for a new Ethernet media system to operate at a speed other than 10x faster than the previous standard. However, the 40 Gb/s speed was added to the 100 Gb/s standards effort to address a concern about the adoption rates for Ethernet technology. Early in the standardization process, a presentation was made to the 802.3ba task force to lobby for the addition of 40 Gb/s. The core of the argument was that while a 10x evolution in speed had worked well for the older 10 Mb/s, 100 Mb/s, and 1 Gb/s Ethernet systems, the data showed that the adoption rate for the 10 Gb/s system had been much slower than for previous systems.

The major reason given for the slower adoption rate was that servers had not been able to support the 10 Gb/s speed for several years after the standard was first developed, leaving only a relatively low-volume market of switch-to-switch interconnections for vendors to sell into. When it came to developing a new 100 Gb/s Ethernet standard, the predictions in 2007 were that server bus and packet processing speeds would not be able to handle a 100 Gb/s speed until roughly 2014 at the earliest.

These considerations led the IEEE to modify the new standards effort in 2007 to include both 40 and 100 Gb/s speeds. The goal was to produce a standard in 2010 that would more closely match the expected performance of servers when the standard was published, and therefore increase the rate at which the new technology would be adopted. The expectation was that a higher-volume server market for 40 Gb/s connections would trigger a "virtuous cycle," with increased sales helping to reduce costs and thus increase the adoption rate even more. By providing 40 Gb/s connections for servers along with

the 100 Gb/s standard, a volume market could be developed years before the same process would occur if the standard only supported 100 Gb/s Ethernet.

The 40 and 100 Gb/s Ethernet systems were both developed in the same 802.3ba supplement that was published in 2010, and both share the same basic architecture. This chapter will describe the 40 Gb/s media types, and Chapter 13 describes the 100 Gb/s media types.

Architecture of 40 Gb/s Ethernet

The 40 Gb/s media system defines a physical layer (PHY) that is composed of a set of IEEE sublayers. Figure 12-1 shows the sublayers involved in the PHY. The standard defines an XLGMII logical interface, using the Roman numerals XL to indicate 40 Gb/s. This interface includes a 64-bit-wide path over which frame data bits are sent to the PCS. The FEC and Auto-Negotiation sublayers may or may not be used, depending on the media type involved.

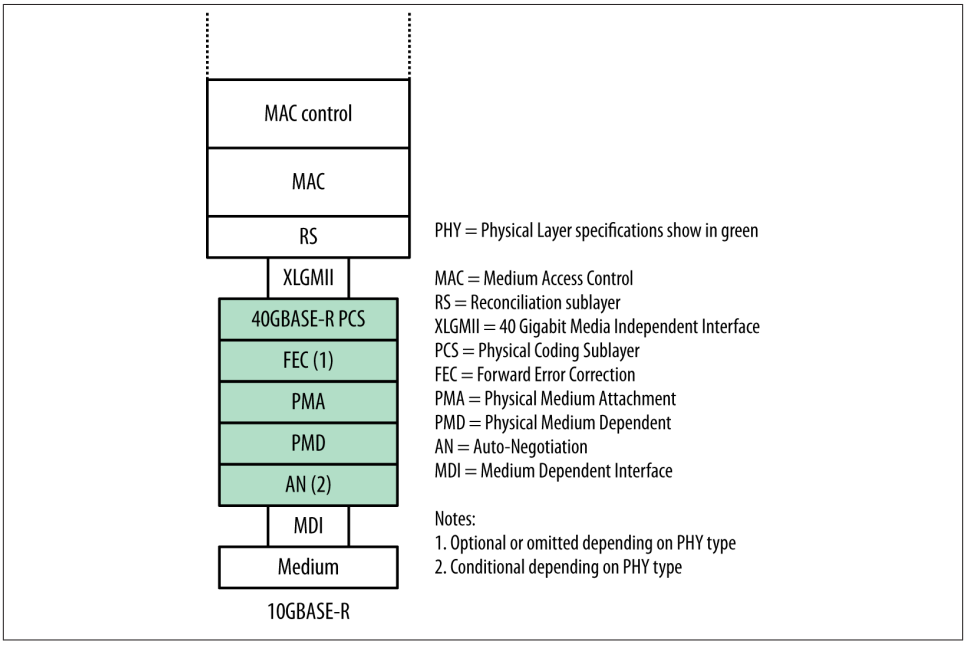


Figure 12-1. 40 Gb/s sublayers

PCS Lanes

To help meet the engineering challenges of providing 40 Gb/s data flows, the IEEE engineers provided a multilane distribution system for data through the PCS sublayer

of the Ethernet interface. The 40 Gb/s speed is challenging, given the state of the art for carrying electrical signals across switch backplanes and across the traces on printed circuit cards.

It's only recently that electrical signals could reach 25 Gb/s on high-volume chip and card interfaces—the set of Common Electrical I/O (CEI) specifications that can provide up to 25 Gb/s for electrical signaling was first published by the Optical Internetworking Forum in 2011.[1] The 40 Gb/s speed is also a challenge when it comes to optical signaling over fiber optic cables, given the cost of the lasers required to provide very rapid signaling.

An important goal for the IEEE engineers was to provide a system that could operate at both 40 and 100 billion bits per second. Another important goal was to provide those speeds using technology that could be implemented at a reasonable cost, and that could be expected to drop in cost as sales volumes increased. These goals were met by reusing technology from the 10 Gb/s standard, and by developing a multilane distribution system that is adaptable as technology changes.

The decision was made to reuse the 64/66-bit line code from 10 Gigabit Ethernet, in which 64 bits of data are transmitted as a 66-bit word after 2 bits are added for identification.

Figure 12-2 shows a schematic diagram of the physical coding sublayer for 10 Gb/s Ethernet, which produces a single PCS lane of data. For the 40 Gb/s standard, the 802.3ba task force developed multilane distribution, which transmits each 66-bit word in a round-robin fashion across four lanes. The Ethernet frame data is broken into 66-bit words, and those words are transmitted simultaneously across all four lanes, with the signaling on each lane operating at a rate of 10.3125 Gbaud.

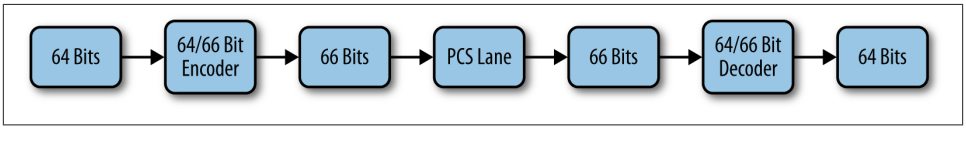


Figure 12-2. PCS lane for 10 Gb/s Ethernet

Figure 12-3 shows the 40 Gb/s multilane system. The Ethernet frame data is encoded as a set of 64-bit chunks of data, each with a 2-bit ID, producing a stream of 66-bit words. The PCS distributes these words across a set of four lanes, in a round-robin fashion. In this example, the four lanes of data simultaneously cross the transmission medium, on four separate fiber optic links. At the receiver, the four streams of data are reassembled

1. Optical Internetworking Forum, "Common Electrical I/O (CEI) - Electrical and Jitter Interoperability agreements for 6G+ bps, 11G+ bps and 25G+ bps I/O," September 1, 2011.

into a correctly sequenced stream of 66-bit words. The words are then reassembled into the frame data, which is handed to the MAC layer as an Ethernet frame.

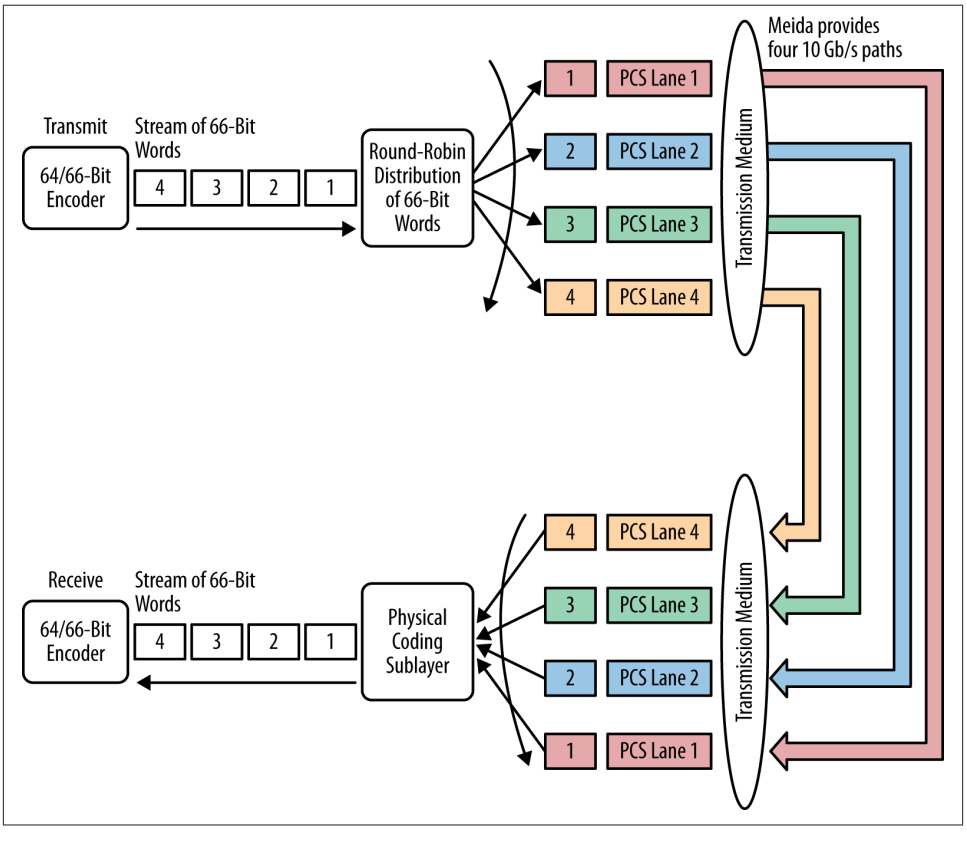


Figure 12-3. PCS multilane for 40 Gb/s Ethernet

PCS lane design and operation

The mapping of the PCS lanes to the physical medium channels, whether electrical ("copper") or optical, is complicated by the fact that the PCS lanes and the physical media systems are separate systems. Advancements in technology for the chip interfaces used to carry PCS data are one thing, while advancements in the copper and optical interfaces used in the media systems are another. The evolution of these different technologies means that PCS lanes and internal printed circuit traces and interfaces may operate at different speeds than the copper and optical interfaces.

To ensure that the design of 40 Gb/s Ethernet allows these technologies to evolve at their own rates, the number of PCS lanes is chosen to provide the least common multiple of the expected capabilities of optical and electrical interfaces. In other words, the number

of lanes provided is expected to be just enough to accommodate the evolution of media types. For 40 Gb/s Ethernet, four PCS lanes, each operating at 10.3125 Gbaud, were defined. The PCS lanes can be combined in various ways to support one, two, and four electrical or optical media channels, depending on the electrical or optical media technology supported. The 100 Gb/s system operates the same way but is provided with more lanes, as described in Chapter 13.

The IEEE standard 40 Gb/s media systems currently use four lanes, based on technology previously developed to support the 10 Gb/s Ethernet standard, and 10 Gb/s optical transmitters and receivers that are widely available at reasonable cost.

However, as faster optical transmitters and receivers become available at reasonable cost, the four PCS lanes could be combined by multiplexing the bits from the four streams into two 20 Gb/s streams, or even a single 40 Gb/s optical stream of data. In the case of two streams, both streams would be transmitted over a media system capable of handling the 20 Gb/s signaling rate per stream. Two 20 Gb/s streams are needed to transmit the combined data of four PCS lanes, as shown in Figure 12-4.

In operation, each PCS lane is provided with lane alignment markers, which are periodically inserted into the flow of bits. The extra bandwidth needed for the alignment markers is created by periodically deleting the interpacket gap (IPG) between Ethernet frames. All multiplexing is done at the bit level, and all of the bits from the same PCS lane follow the same physical path.

The PCS alignment marker also provides information for the deskew operation, in which the receiver removes the markers and realigns the lanes of data to compensate for any signal speed variance, or "skew," that may have occurred during transmission over the media system. Any difference in rate resulting from the deleted markers is compensated for at the receiver by inserting IDLE symbols to maintain the correct timing.

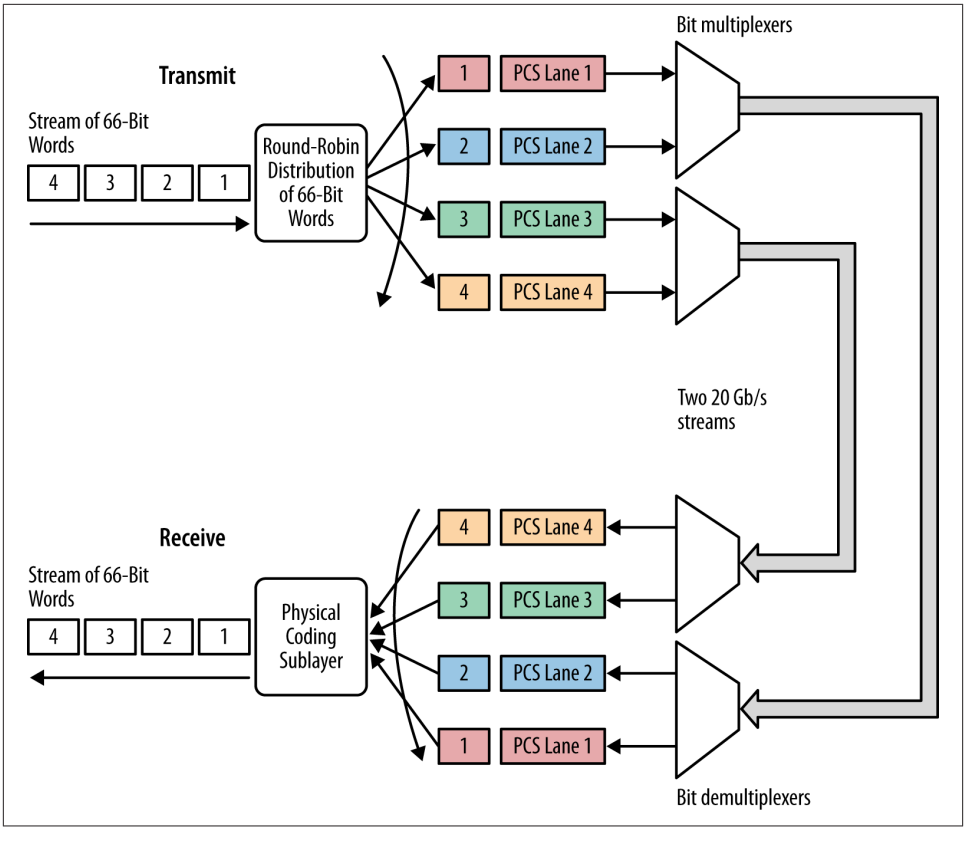


Figure 12-4. PCS lanes over a faster media system

Multiple PCS lanes are not aggregated links

You should not confuse the system of multiple PCS lanes with the operations of an Ethernet aggregated link as defined in the 802.3ad/802.1AX link aggregation standard.

Link aggregation was originally standardized in 2000, in the 802.3ad supplement. In 2008, link aggregation was moved to the 802.1 standard in the 802.1AX-2008 supplement.

The multilane PCS distribution method of carrying Ethernet frame data across four lanes avoids the constraints of a link aggregation channel composed of four individual 10 Gb/s link segments. In a link aggregated channel providing 40 Gb/s over four 10 Gb/s links, the 802.1AX standard specifies that a given flow of data is always transmitted over

a single 10 Gb/s link in the channel. Therefore, no single flow of data between two stations can exceed 10 Gb/s across the aggregated channel.

In 40 Gb/s Ethernet, a single flow of data is transmitted over the link segment by simultaneously striping the data across four PCS lanes, each operating at 10.3125 Gbaud. The end result is true 40 Gb/s throughput for a single flow of data, and for all frames sent over the link.

40 Gigabit Ethernet Twisted-Pair Media Systems (40GBASE-T)

In July 2012, a "Call for Interest" was issued for "Next Generation BASE-T," to gauge the interest level in producing a twisted-pair standard beyond 10 Gb/s and to evaluate the technical challenges. Engineering analysis found that balanced twisted-pair cabling could carry 40 Gb/s signals for a useful distance, making a new 40 Gb/s twisted-pair standard possible.

The value of providing a 40GBASE-T connection includes multispeed Auto-Negotiation support over the RJ45-style connector, lower cost, and, due to the smaller size of the RJ45 style connector, an ability to provide this media type as a built-in interface on server motherboards. The market analysis found that when interfaces come built-in and "for free" on servers, it leads to much faster adoption of the technology.

A 40GBASE-T standard will also make it possible for vendors to provide a built-in twisted-pair server interface that works at multiple speeds, making the adoption of 40 Gb/s Ethernet an easy process. New servers with multispeed interfaces supporting 1, 10, and 40 Gb/s operation will be able to continue to operate at the currently supported speed of the network connection. When the network is upgraded to 40 Gb/s, the server interface can use Auto-Negotiation to automatically upgrade its network interface speed to 40 Gb/s.

An IEEE Project Authorization Request for a 40GBASE-T task force was approved on September 25, 2012, and the task force has begun work on a new 802.3bq 40GBASE-T supplement to the standard. The objective of 802.3bq is to provide 40 Gb/s operation over a balanced twisted-pair segment that is 30 m in length, with up to two connectors, and to support Auto-Negotiation and Energy Efficient Ethernet.

Providing 10GBASE-T operation over Cat6A twisted-pair cabling was a significant engineering challenge, requiring improved cabling specifications to support 400 MHz signaling over the cabling. Providing 40GBASE-T operation over twisted-pair cable is an even greater challenge, and quite difficult to do. One way to achieve better signal-handling capability over twisted-pair cabling is to use a shorter segment length, which is why there is a 30 m target length in the 40GBASE-T standard. Even with a shorter segment length, the 40GBASE-T standard also requires a new cabling specification to

handle the increased signaling rates needed. The new specifications for "Category 8" cabling designed to support 40GBASE-T signals are currently under development in the ANSI/TIA 568 C.2-1 Category 8 standardization project.

Progress on the new 40GBASE-T standard will depend on how quickly the technological challenges can be identified and resolved. Depending on the rate of progress, it is possible that formal adoption of the new standard could take place in late 2015 or early 2016.

40 Gigabit Ethernet Short Copper Cable Media Systems (40GBASE-CR4)

The 40GBASE-CR4 short reach copper segment is defined in Clause 85 of the standard. It specifies a media system based on four lanes of PCS data carried over four twinaxial cables. A twinaxial cable is similar to coaxial cable, except that each twinaxial cable has two inner conductors instead of the single conductor found in coaxial cable. Twinaxial cable is capable of carrying high-speed signals for a relatively short distance. The standard specifies a segment length of up to 7 m.

The 40GBASE-CR4 standard defines a medium dependent interface that is based on a *quad small form-factor pluggable* (QSFP+) connector, referenced in the IEEE standard as small form-factor specification SFF-8436. The QSFP+ module is not standardized by a formal standards group, but instead is specified by a multisource agreement (MSA) developed by competing manufacturers.[2]

Multisource agreements are the primary method used these days to develop communications connectors and transceiver modules for Ethernet and other network systems. As technology advances, cabling and equipment vendors work together to develop smaller and more efficient connectors and modules, using an MSA as a rapid method of developing these enhancements in a standardized way that will provide interoperability among equipment from different vendors.

While the segment length is specified up to 7 m, some vendors provide both active and passive versions of the 40GBASE-CR4 cable, with the active versions capable of longer segment lengths. For example, a vendor may offer lengths of 1, 3, and 5 m in passive cables, with 7 and 15 m supported in active cables. You will find variations in supported cable types and lengths among different vendors, and it's up to you to verify the supported cable lengths on the equipment you purchase.

The QSFP+ transceiver and connector module used on the copper cables is the same basic module as that used on 40 Gb/s optical fiber links. However, rather than providing

2. The SFF-8436 specifications have been replaced by the "INF-8438 Specification for QSFP (Quad Small Formfactor Pluggable) Transceiver."

an optical transceiver at each end to which a fiber optic cable is connected, the 40GBASE-CR4 cable uses the QSFP+ module but leaves out the expensive optical lasers.

Figure 12-5 shows a fixed-length 40GBASE-CR4 direct attach cable segment, which is always sold with the QSFP+ modules permanently attached to each end of the cable. This cable provides four pairs of conductors in a fairly thick cable, with an outer diameter ranging from 6.1 mm (0.24 inches) for a 1 m cable up to 9.8 mm (0.39 inches) for a 7 m cable length. The bend radius is typically specified as being 10 times the outer diameter. That would lead to a bend radius of from 6.1 cm (2.4 inches) to 9.8 cm (3.85 inches) for the cable outer diameters mentioned.

Figure 12-5. 40GBASE-CR4 QSFP+ direct attach cable

The QSFP+ module itself is roughly three inches in length, not including the plastic tab that, when pulled, disengages the connector from the port. This makes for a fairly long connector assembly on each end of the cable, and to connect anything other than closely associated equipment, you must be able to route the cable and the permanently attached QSFP+ modules through any intervening cable management trays and cable guides.

While the QSFP+ transceiver module has 38 contacts on it, the 40GBASE-CR4 standard only specifies the set of contacts needed for transmitting and receiving four lanes of data. The signal crossover from source lane (Tx) to destination lane (Rx) is provided by the wiring scheme specified in the standard.

Table 12-1 lists the QSFP+ module contact positions used for 40 Gb/s operation. These contacts support the four source lanes (SL 0 through 3) and four destination lanes (DL 0 through 3), with a positive and negative wire for each lane to support the differential signaling. Multiple signal grounds are provided to help maintain signal quality.

Table 12-1. 40GBASE-CR4 signals and QSFP+ contact positions

Tx lane	Contact	Rx lane	Contact
Signal GND	S1	Signal GND	S13
SL1<neg>	S2	DL2<pos>	S14

Tx lane	Contact	Rx lane	Contact
SL1<pos>	S3	DL2<neg>	S15
Signal GND	S4	Signal GND	S16
SL3<neg>	S5	DL0<pos>	S17
SL3<pos>	S6	DL0<neg>	S18
Signal GND	S7	Signal GND	S9
SL2<pos>	S33	DL1<neg>	S21
SL2<neg>	S34	DL1<pos>	S22
Signal GND	S35	Signal GND	S23
SL0<pos>	S36	DL3<neg>	S24
SL0<neg>	S37	DL3<pos>	S25
Signal GND	S38	Signal GND	S26

40GBASE-CR4 Signaling Components

A QSFP+ 40GBASE-CR4 port on a switch or Ethernet interface may support either active or passive direct attach cables, or both. Given that the QSFP+ transceiver module is used for both copper and optical cables, a QSFP+ port may also support either copper or optical modules, or both. The vendor chooses which of these media types to support on a QSFP+ port, and you must refer to the documentation for the interface or switch port to determine what is supported.

When a QSFP+ port supports a 40GBASE-CR4 connection, all you need to do is insert the QSFP+ module on the end of the cable into the port until it latches. The QSFP+ active and passive cable assemblies are hot pluggable, meaning that you can safely connect and disconnect them with power enabled on the switch or computer interface.

Figure 12-6 shows a set of QSFP+ ports on a large Ethernet switch, a QSFP+ direct attach cable, and an Ethernet interface with two QSFP+ DA ports. The twinaxial cable has permanently attached QSFP+ modules on each end, and the QSFP+ modules are plugged into the QSFP+ ports on the switch and the Ethernet interface.

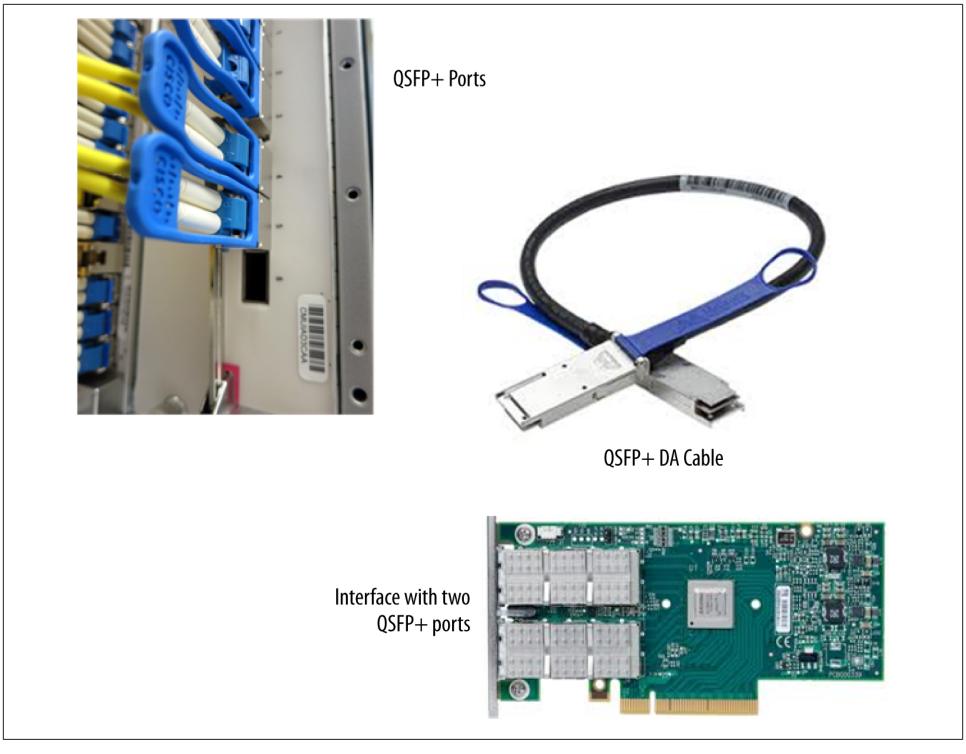


Figure 12-6. 40GBASE-CR4 connection components

40GBASE-CR4 Signal Encoding

The direct attach cables and their QSFP+ connector modules use electrical signaling defined in the standard as a "low-swing AC coupled differential interface." This type of differential signaling provides noise immunity and reduced levels of electromagnetic interference. The differential signal results in a voltage on the wire that is roughly two volts, peak to peak. There are 4 pairs of conductors carrying signals in each direction, for a total of 8 pairs of conductors, or 16 wires, in the cable segment.

The 40GBASE-CR4 link transfers four lanes of encoded and scrambled data at 10.3125 Gbaud, to accommodate both data and the overhead associated with 64B/66B coding. The self-clocked nature of the signal eliminates skewing between clock and data signals. The electrical interface to the media is based on 100 ohm cable impedance, while the signal termination electronics provide both differential and common-mode suppression of signal noise and reflections. The direct attach copper cable is designed to meet a bit error ratio specification of 10^{-12}, which is a potential for 1 bit error in every trillion bits sent.

QSFP+ Connectors and Multiple 10 Gb/s Interfaces

A vendor can design a 40 Gb/s Ethernet interface and its associated QSFP+ port in such a way that the port can provide either a 40 Gb/s Ethernet interface, or four independent 10 Gb/s Ethernet interfaces. As we've seen, each lane of the 40 Gb/s Multilane PCS signaling is identical to the single 64/66B encoded PCS lane that is used in the 10 Gb/s standard.

This makes it possible for a vendor to design a 40 Gb/s Ethernet interface to allow configuration of the internal signaling paths so as to provide either a single 40 Gb/s Ethernet interface, or four individual 10 Gb/s Ethernet interfaces. Note that this does not mean that a 40 Gb/s interface is simply operating as four 10 Gb/s interfaces. Instead, it means that the internal signaling paths can be configured to provide four PCS lanes of data that combine to operate as a single 40 Gb/s interface, or four individual 10 Gb/s interfaces, each using a single PCS lane. While a vendor can support the ability to configure a QSFP+ port to provide four 10 Gb/s interfaces, you cannot assume that every 40 Gb/s QSFP+ port can do this. Once again, you must refer to the documentation to determine what is supported.

As shown in Figure 12-7, this connection uses a four-to-one cable, also called a "breakout cable," to break out the single QSFP+ port into four direct attach cables with SFP+ connectors permanently attached to their ends. Each of the SFP+ connectors is a separate 10 Gb/s direct attach Ethernet transceiver, which can be plugged into an SFP+ direct attach port. This supports a short distance connection. One vendor, for example, supports 1, 3, and 5 m lengths using passive cables, and 7 and 10 m lengths with active cables.

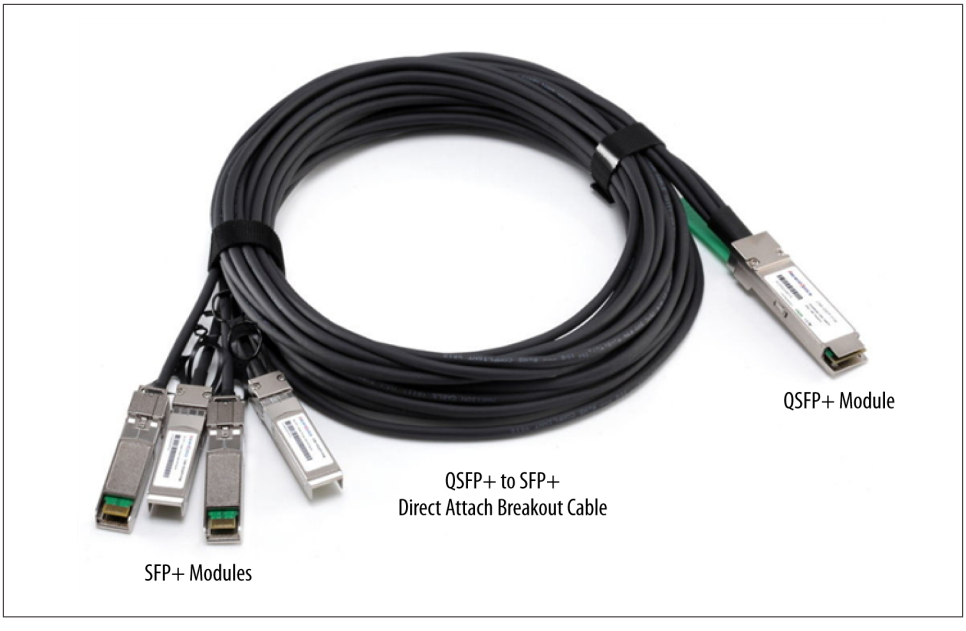


Figure 12-7. QSFP+ breakout cable

This flexibility makes it possible for vendors to provide QSFP+ ports on a switch that can support either 40 Gb/s connections, or four 10 Gb/s connections. Of course, you will still need to route the direct attach cable with its four SFP+ connectors through any cable management system that may be present. This can be a challenge when the cable management is already holding cables, making it difficult to route a bulky cable and its four permanently attached SFP+ connectors through the cabling system.

40 Gigabit Ethernet Fiber Optic Media Systems

There are two 40 Gigabit fiber optic physical medium dependent (PMD) specifications in the standard, which provide 40 Gb/s Ethernet over multimode fiber optic (MMF) cable and single-mode fiber optic (SMF) cable. The 40GBASE-SR4 short reach fiber optic system sends four lanes of PCS data over four pairs of multimode cables, for a total of eight fiber strands. The 40GBASE-LR4 long-reach system sends four lanes of PCS data over four wavelengths of light, carried over a single pair of fiber optic cables.

The first 40 Gb/s transceivers were based on the *C form-factor pluggable* (CFP) module, which is a large module capable of handling up to 24 watts of power dissipation. First-

generation transceivers with multiple chips and larger power requirements were based on this module. The CFP module is specified by a multiagreement.[3]

Figure 12-8 shows a CFP module, which can be used to provide either a 40GBASE-SR4 or a 40GBASE-LR4 transceiver. The module shown in this figure is a 40GBASE-LR4 transceiver, which provides two SC fiber optic connectors for connection to a pair of single-mode fibers. The operation of the 40GBASE-LR4 connection is described later in this chapter.

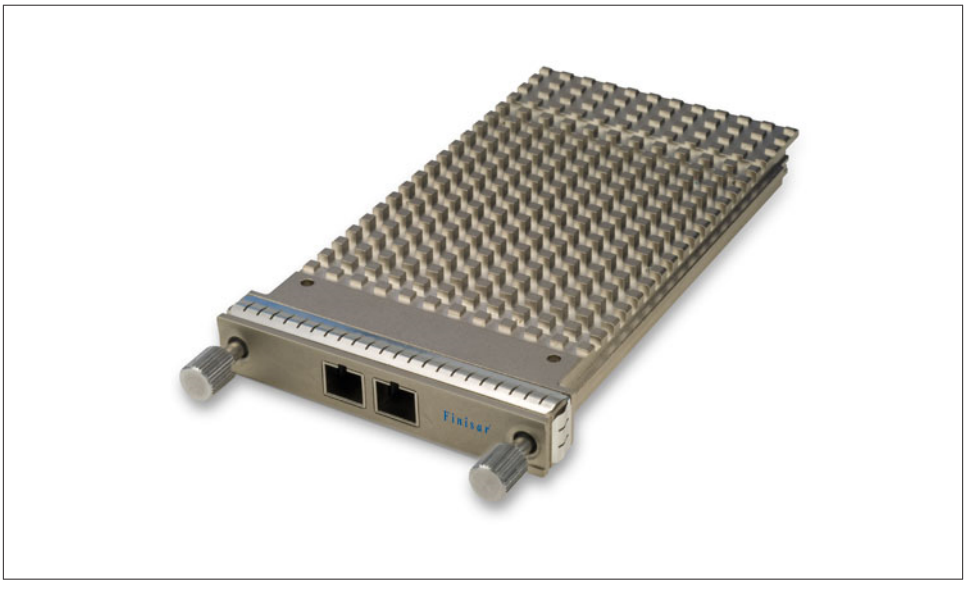


Figure 12-8. 40 Gigabit CFP transceiver module

The most popular connector for 40 Gb/s interfaces these days is the QSFP+ module: it takes up much less space on a switch or server interface, making it possible for vendors to provide multiple QSFP+ ports in the same space as a single CFP port. The QSFP+ transceiver module for 40GBASE-SR4 is provided with a multifiber push-on (MPO) media connector, carrying multiple pairs of fiber optic cables to support the four lanes of data for the short reach fiber standard. The 40GBASE-LR4 long-reach system uses a QSFP+ transceiver equipped with a duplex fiber connector for connecting to the single pair of fiber cables.

Figure 12-9 shows an MPO plug connector on a multimode fiber cable that supports 12 individual fibers (6 pairs). This connection supports the eight fiber cables needed to provide a 40GBASE-SR4 connection, plus four fibers that are unused. The MPO plug

3. Revision 1.4 of the "CFP MSA Hardware Specification."

has two alignment pins, which help keep each connector and its fibers correctly aligned when the two connectors are mated. Also shown is an end-on view of a plug connector, which shows the 12 fibers lined up between the two alignment pins. Note that the key on the plug connector ensures that the plug connector will be correctly oriented in the socket.

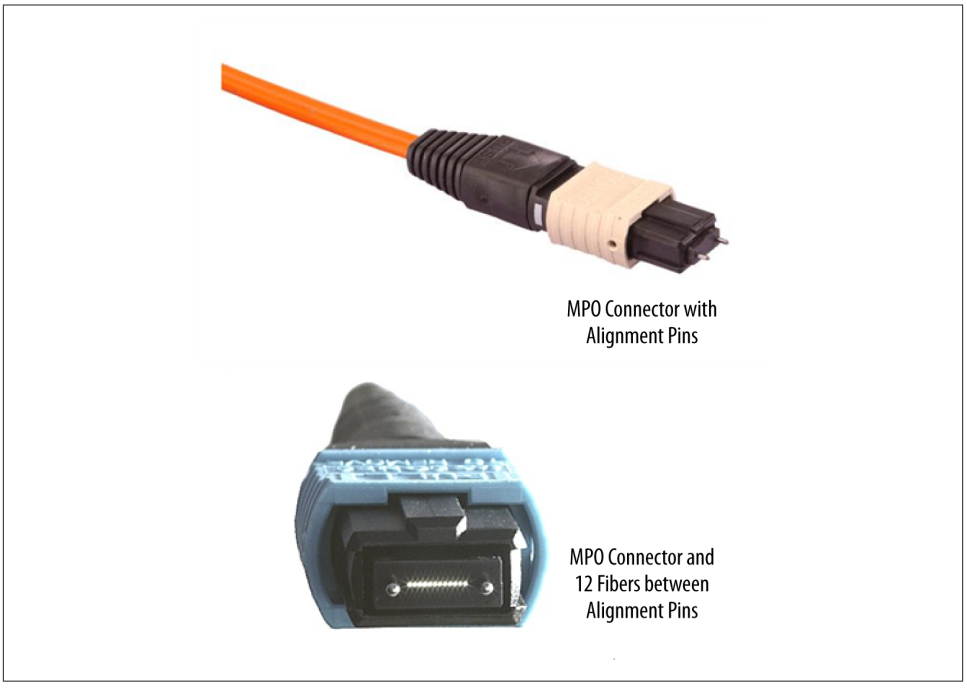


Figure 12-9. 40 Gigabit multimode QSFP+ transceiver module and connectors

Figure 12-10 shows the TX and RX connections specified in the standard for an MPO cable with 12 fibers. Only 8 of the 12 fibers are required for 40GBASE-SR4 operation, leaving 4 of the fiber optic strands in a 12-fiber MPO ribbon cable unused. More information on MPO cables and connectors can be found in Chapter 17.

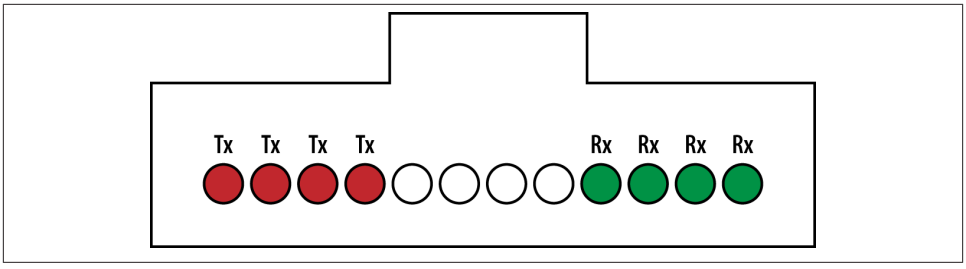
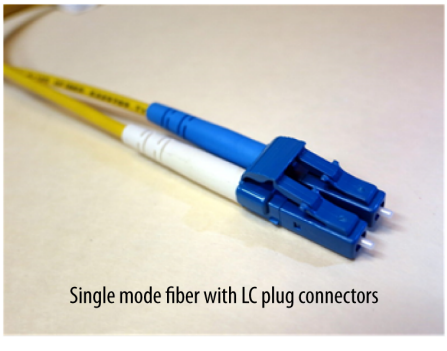


Figure 12-10. MPO connections for 40GBASE-SR4

Figure 12-11 shows a single-mode fiber optic cable terminated in an LC optical plug connector. Also shown is a QSFP+ transceiver module with LC optical sockets.


Single mode fiber with LC plug connectors


QSFP+ transceiver module with LC socket connectors

Figure 12-11. 40 Gigabit QSFP+ transceiver for single-mode fiber

Optical cables can be ordered in specific lengths and with specific connector types, as required by your circumstances. If you are connecting an optical jumper cable from a QSFP+ module in a switch port directly to a server interface, then you need a short cable

with the correct cable connectors on each end. For 40GBASE-SR4, this cable would have MPO connectors on each end. If you are connecting a QSFP+ module in a switch port to an optical connector termination point in your data center or cabling closet, the optical connectors in your termination point might be different than the ones used on the 40 Gb/s transceiver module, requiring a cable with a different type of optical connector on each end.

Maximum segment lengths are dependent on a number of factors. Fiber optic segment lengths will vary depending on the cable type and wavelength used by the media type. More information on multimode and single-mode fiber optic segments and components can be found in Chapter 17.

40 Gb/s Fiber Optic Media Specifications

The 40 Gbit standard provides specifications that the multimode and single-mode fiber optic media must meet to support 40 Gigabit Ethernet.

40GBASE-SR4 media specifications

The 40GBASE-SR4 media type is based on multimode fiber optic cables. Multimode fiber optic components are less expensive than single-mode components and are able to transmit light for relatively short distances compared to single-mode systems.

Table 12-2 shows the distances and channel insertion losses for 40GBASE-SR4. The modal bandwidth of a multimode fiber optic cable refers to its signal-carrying characteristics. Higher modal bandwidth means that the cable can transmit a signal for a longer distance while maintaining sufficient signal quality to the optical receiver in the Ethernet interface at the end of the link. The 50 µm refers to the diameter of the signal-carrying portion of the fiber optic cable.

Table 12-2. Optical specifications for 40GBASE-SR4

Fiber type	Minimum modal bandwidth at 850 nm (MHz-km)	Channel insertion loss (dB)	Operating range (m)
50 µm MMF (OM3)	2,000	1.9	0.5 to 100
50 µm MMF (OM4)	4,700	1.5	0.5 to 150

You may wonder why the maximum media lengths for 40GBASE-SR4 are shorter than the lengths specified in the 10GBASE-SR media system (see Table 11-5). Given that the 10GBASE-SR maximum segment lengths are 300 m and 400 m over OM3 and OM4 fiber, respectively, and given that the 40GBASE-SR4 lanes are based on the same technology used in 10GBASE-SR, why can't the 40GBASE-SR4 system provide the same maximum segment lengths?

The answer is that the transceiver specifications were changed in 40GBASE-SR4 to allow for less-expensive optical components to be used, given that there are four transmitters

and receivers in each transceiver. The major changes were to relax some of the timing requirements, and to relax certain optical transmission specifications. The maximum segment length was reduced, in turn, to maintain signal quality when using the lower-cost optical transmitters and receivers.

The total link segment optical power budget for both OM3 and OM4 fiber types is 8.3 dB, with varying amounts of power consumed by optical noise characteristics, intersymbol interference issues, and the like, depending on the type of the multimode fiber. The channel insertion loss is that portion of the link segment optical power budget that can be consumed by the optical cabling and connectors used on a given segment. As long as the optical power loss, as measured through the link segment from end to end, is at or below the level shown for the channel insertion loss, then the segment will operate correctly.

40GBASE-LR4 media specifications

The 40GBASE-LR4 media system is designed to operate over single-mode fiber optic cables, with the specifications shown in Table 12-3. The specifications for the long-reach optical fibers used in 40GBASE-LR4 are simpler, because the transmission characteristics for single-mode fiber optic cables do not include the modal bandwidth considerations found in multimode fiber specifications.

Table 12-3. Optical specifications for 40GBASE-LR4

Optical type	Channel insertion loss (dB)	Operating distance
9 µm SMF	6.7	2 m to 10 km

The total link power budget for the 40GBASE-LR4 system is 9.3 dB, with varying amounts of power consumed by optical noise characteristics, intersymbol interference issues, and the like. The channel insertion loss is that portion of the link segment optical power budget that can be consumed by the optical cabling and connectors used on a given segment.

Vendor-specific short-range media specifications

At least one major vendor provides a version of the 40GBASE-SR4 transceiver designed to work over a wider range of multimode fiber optic cables. The goal is to support 40 Gb/s operation over installed fiber optic systems with older cabling types that do not have the higher performance of OM3 and OM4 cables. Cisco Systems provides a version of QSFP module for 40 Gb/s operation over multimode fiber that it identifies as QSFP-40G-CSR4.

As shown in Table 12-4, the Cisco Systems QSFP-40G-CSR4 module works over a wider variety of fiber and with different fiber distances than the fiber specified in the standard. Because this is a vendor-specific media type, both ends of the link must be using equipment from the same vendor to ensure correct operation.

Table 12-4. Optical specifications for Cisco's QSFP-40G-CSR4 module

Fiber type	Minimum modal bandwidth at 850 nm (MHz-km)	Channel insertion loss (dB)	Operating range (m)
62.5 µm MMF OM1	200	2.6	0.5 to 33 m
50 µm MMF OM2	500	2.6	0.5 to 82 m
50 µm MMF OM3	2,000	2.6	0.5 to 300
50 µm MMF OM4	4,700	2.9	0.5 to 400

Vendor-specific bidirectional short-range optical transceivers

An example of the kind of vendor-specific innovation that occurs in the Ethernet market is the bidirectional short-range optical transceiver, developed by Cisco Systems. This transceiver provides 40 Gb/s operation over a single pair of fiber optic cables, by providing two 20 Gb/s optical channels on two separate wavelengths (850 and 900 nm) over each multimode fiber optic strand.

The result is a 40 Gb/s link segment that can operate over the same duplex fiber path that previously supported a 10 Gb/s link segment, making the upgrade from 10 to 40 Gb/s easier to achieve. The maximum segment length achievable is 100 m (328 feet) over OM3 and OM4 fiber.[4]

Implementing this solution requires a Cisco 40 Gb/s bidirectional transceiver that is supported only on Cisco switch ports. Given that servers and switches from other vendors do not support this transceiver, this technology is limited to uplink connections between high performance Cisco switches that support this media type, such as those found in data centers.

40GBASE-LR4 Wavelengths

The four lanes of PCS data are sent over four wavelengths of light in the 40GBASE-LR4 long reach over single-mode media system. All four wavelengths of light are sent over a single pair of fiber optic cables, using a system called *coarse wave division multiplexing* (CWDM). Each wavelength of light has a specific CWDM frequency, as specified in the *wavelength grid* definition in the ITU-T G.694.2 standard.

Table 12-5 shows the four center wavelengths, and the frequency range for each of the four wavelengths, or "colors," of light used to carry signals over the single pair of single-mode cables. Each 40GBASE-SR4 transceiver contains a four-wavelength optical transmitter and a four-wavelength receiver. The single-mode optical cable simply carries the four wavelengths of light between the transceivers, providing four paths for the four lanes of PCS data, each signaling at the rate of 10.3125 Gbaud per second. As we've

4. Cisco published a data sheet for this transceiver, as well as a white paper on its use and data center cabling approaches.

described, the four PCS lanes carry the Ethernet frame data between the interfaces at 40 Gb/s.

Table 12-5. 40GBASE-LR4 wavelengths

Lane	Center wavelength	Wavelength range
L_0_	1,271 nm	1,264.5 to 1,277.5 nm
L_1_	1,291 nm	1,284.5 to 1,297.5 nm
L_2_	1,311 nm	1,304.5 to 1,317.5 nm
L_3_	1,331 nm	1,324.5 to 1,337.5 nm

40 Gigabit Extended Range

The 40 Gigabit Ethernet standard published in 2010 did not include an extended-range media type for 40 Gb/s operation. However, a new supplement to the standard is being developed that includes the goal of providing a 40GBASE-ER4 media type for extended-range operation over single-mode fiber optic cables.

The Project Authorization Request for the 802.3bm supplement was granted in May 2012, and work has been proceeding since then. Assuming that good progress is made, the specifications may be complete sometime in late 2014, with formal adoption by the standards organization expected in 2015.

CHAPTER 13

100 Gigabit Ethernet

The development of 100 Gb/s Ethernet began with a Higher Speed Study Group "Call For Interest" meeting in July 2006. In response, an IEEE task force was formed to develop 100 Gb/s Ethernet specifications in the 802.3ba supplement. As described in the previous chapter, this effort was expanded to include 40 Gb/s Ethernet, and the combined 100 and 40 Gb/s 802.3ba supplement was completed and published in 2010.

The 40 and 100 Gb/s Ethernet systems were developed together and share the same basic architecture. This chapter describes the 100 Gb/s media types.

Architecture of 100 Gb/s Ethernet

The 100 Gb/s media system defines a physical layer (PHY) that is composed of a set of IEEE sublayers. Figure 13-1 shows the sublayers involved in the PHY. The standard defines a CGMII logical interface, using the Roman numeral C to indicate 100 Gb/s. This interface defines a 64-bit-wide path, over which frame data bits are sent to the PCS. The FEC and AN sublayers may or may not be used, depending on the media type involved.

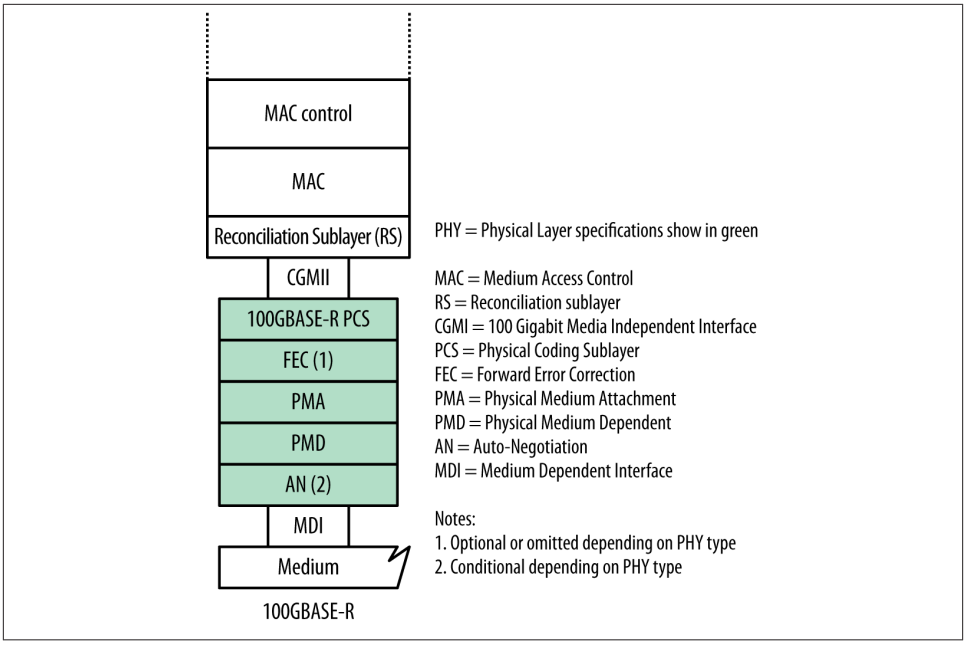


Figure 13-1. 100 Gb/s sublayers

PCS Lanes

One important goal for the IEEE engineers working on the 802.3ba standard was to provide a system that could support both 40 and 100 billion bits per second. An equally vital goal was to implement those new Ethernet speeds with a technology that could be developed at a reasonable expense, and that could be expected to drop in cost as sales volumes increased. These goals were met by reusing technology from the 10 Gb/s standard, as well as by developing a multilane distribution system for the PCS sublayer that is adaptable as technology changes.

PCS lane design and operation

The operation of the multilane PCS system is described in more detail in Chapter 12. The number of PCS lanes for the 100 Gb/s system was chosen to provide enough lanes to accommodate the evolution of internal interface technology and media types. For 100 Gb/s Ethernet, 20 PCS lanes were defined.

These 20 PCS lanes can be multiplexed into any of the supported interface widths, depending on the electrical (copper) or optical media technology being deployed. The number of electrical or optical interface widths supported is equivalent to the number of factors of the total set of lanes. Therefore, 20 PCS lanes can support interface widths of 1, 2, 4, 5, 10, and 20 channels or wavelengths.

Figure 13-2 provides a look at how 20 PCS lanes are multiplexed onto a fiber optic medium. In this example, we are assuming that four lanes of data are being transmitted over four wavelengths of light carried over a single pair of fiber optic cables. The 20 PCS lanes carry the Ethernet frame data, which is being transmitted through the lanes as 64-bit chunks of data with a 2-bit header, for a total of 66 bits per block of data.

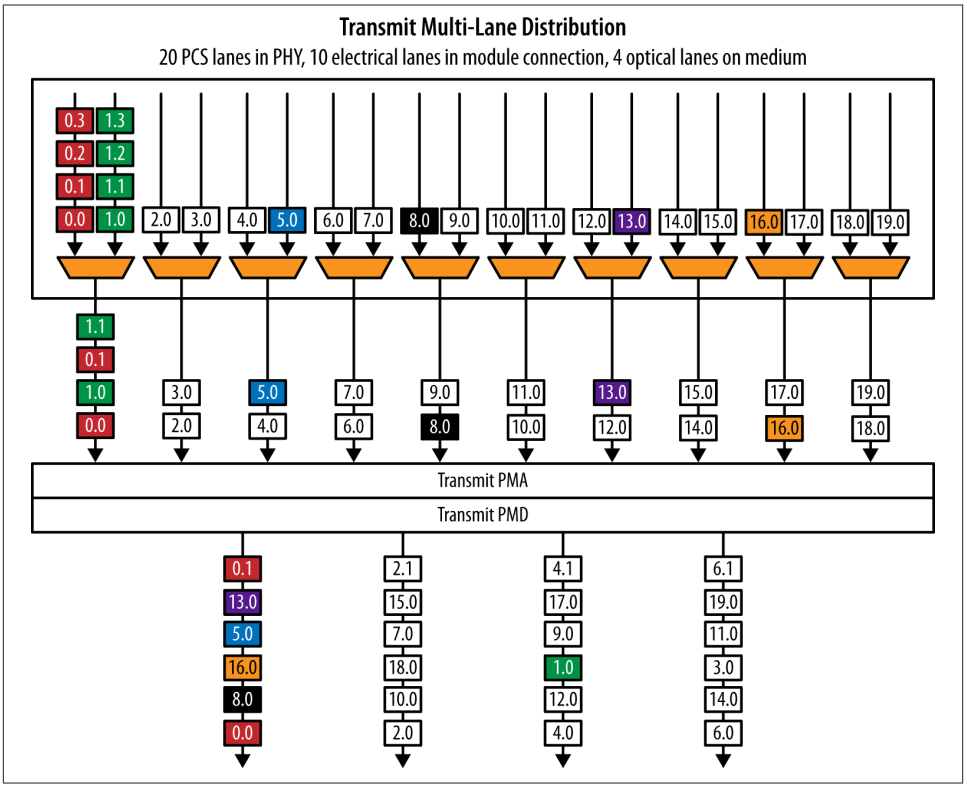


Figure 13-2. 100 Gb/s multilane transmit operation

In operation, each PCS lane is provided with lane alignment markers, which are periodically inserted into the flow of bits. An alignment word is sent on all lanes simultaneously, after every 16,384 blocks. The extra bandwidth needed for the alignment markers is created by deleting interpacket gap (IPG) characters transmitted between Ethernet frames, as needed, while leaving a minimum IPG of one character. The rate adjustment maintains the bit rate through the interface at 100 Gb/s.

All multiplexing is done at the bit level, and all of the bits from the same PCS lane follow the same electrical and optical path. This ensures that data from a lane is received in the correct bit order at the other end of the link.

The PCS lane operation is shown in Figure 13-2 with the frame data blocks in lane zero numbered as 0.0, 0.1, 0.2, the blocks in lane one as 1.0, 1.1, 1.2, and so on. Next, we show the multiplexing function, which uses a round-robin process to copy the 20 lanes of PCS data onto 10 electrical lanes for transmission into the transceiver module. In the following layer of transceiver operation, a second multiplexing process uses the same round-robin process to copy the 10 lanes of data onto 4 lanes for transmission onto the medium, at an effective data rate of 25 Gb/s per lane.

Figure 13-3 shows what happens at the other end of the link, where the four lanes of data are received. A reverse process of demultiplexing takes place, in which the 4 lanes of data received from the optical medium are unpacked onto 10 electrical lanes, and the 10 lanes are then unpacked onto 20 PCS lanes.

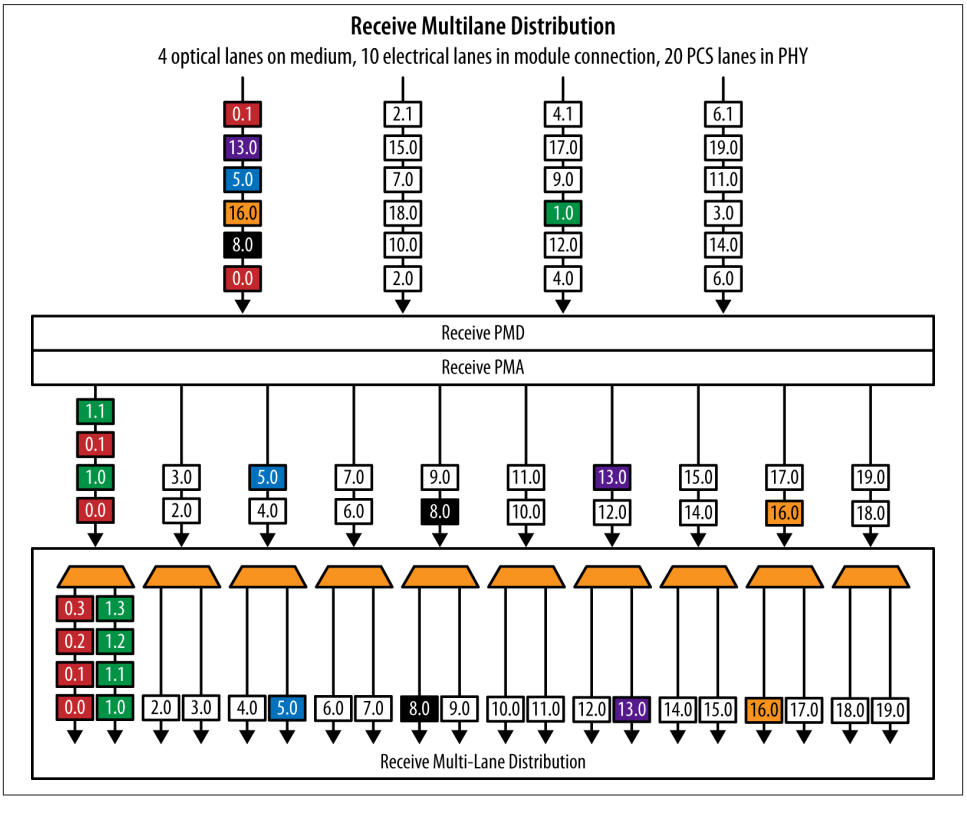


Figure 13-3. 100 Gb/s multi-lane receive operation

During transmission across the medium, data bits being transmitted in one lane may arrive slightly faster than the bits in another lane, depending on factors such as the transmission speed of the medium, the path through the transceiver and interface elec-

tronics, and so on. This effect is known as "skew," and the receiver is responsible for deskewing the received data and retiming it, so that the data in all lanes is received in the correct order.

The PCS alignment markers provide information for the skew compensation operation, in which the receiver removes the alignment markers and realigns the lanes of data to compensate for any signal speed variance. Any difference in bit rate resulting from the deleted markers is compensated for at the receiver by inserting IDLE symbols to maintain the correct timing. The final result is the transmission of frame data across the link, to be received by the Ethernet interface at the other end of the link at a rate of 100 Gb/s.

100 Gigabit Ethernet Twisted-Pair Media Systems

There are no current plans to develop a twisted-pair media system for 100 Gigabit Ethernet. The development of 40 Gb/s Ethernet over twisted-pair cabling is based on a segment length that is limited to 30 m, and to achieve that goal the twisted-pair media will have to support roughly 2 Gbaud of signaling per wire pair. Scaling this up to 100 Gb/s is not currently feasible.

Improved versions of twisted-pair cabling and improved signal processing techniques have increased the high-speed signaling capacity of twisted-pair cabling. However, on the assumption that 100 Gb/s operation might require roughly 20 Gbaud of signaling capability per wire pair, it is not expected that 100 Gb/s operation will be economically achievable. Even it if were possible to produce twisted-pair cabling that could carry signals at that rate for a useful distance, as well as signaling systems at each end of the link that met the performance requirements, it is unlikely that the technology could be produced at a cost that would be successful in the marketplace.

100 Gigabit Ethernet Short Copper Cable Media Systems (100GBASE-CR10)

The 100GBASE-CR10 short reach copper segment is defined in Clause 85 of the standard. It specifies a media segment based on 10 lanes of PCS data carried over 10 twinaxial cables or equivalent cabling that meets the electrical specifications. A twinaxial cable is similar to coaxial cable, except that each twinaxial cable has two inner conductors instead of the single conductor found in coaxial cable. Twinaxial cable is capable of carrying high-speed signals for a relatively short distance. The standard specifies a segment length of up to 7 m.

The 100GBASE-CR10 standard defines a medium dependent interface that is based on a CXP module with a "Mini Multilane" connector. The connector is referenced in the IEEE standard as small form-factor specification SFF-8642, and the CXP module was

initially developed for use in the InfiniBand network system. The CXP module and mini-multilane connector module are not standardized by a formal standards group, but instead are specified by a multisource agreement (MSA) that is developed by competing manufacturers.[1] According to the InfiniBand specification, the CXP name derives from the use of the Roman numeral C for 100, and XP for "eXtended-capability Pluggable form-factor."

Multisource agreements are the primary method used these days to develop communications connectors and transceiver modules for Ethernet and other network systems. As technology advances, cabling and equipment vendors work together to develop smaller and more efficient connectors and modules, using an MSA as a rapid method of developing these enhancements in a standardized way that will provide interoperability among equipment from different vendors.

While the 100BASE-CR10 cable segment is defined in the standard, no vendors are currently providing this short reach segment. The 100 Gb/s Ethernet system is the most recent and highest-speed standard, and current offerings are expensive due to the newness of the technology and the low sales volumes. Given the technical challenges, the marketplace usually first sees the development of fiber optic interfaces for new high-speed Ethernet systems, with the lower-cost copper media systems coming later as technology improves and costs are reduced.

The 100BASE-CR10 CXP transceiver module has 84 contacts, with 48 positions allocated by the InfiniBand CXP specification for differential signals, 28 for signal grounds, 4 for power connections, and 4 for control signals. The 100GBASE-CR10 standard specifies only the set of contacts needed for transmitting and receiving 10 lanes of data. The signal crossover from source lane (Tx) to destination lane (Rx) is provided by the wiring scheme specified in the standard.

Table 13-1 lists the CXP module contact positions used for 100 Gb/s operation. These contacts support the 10 source lanes (SL 0 through 9) and 10 destination lanes (DL 0 through 9), with a positive and negative wire for each lane to support the differential signaling. Multiple signal grounds are provided to help maintain signal quality.

1. See the Rev 2.9 SFF-8642 specifications, "SFF-8642 Specification for Mini Multilane 10 Gb/s 12X Shielded Connector,". The InfiniBand Trade Association (IBTA) published the CXP specification in September 2009 as "Annex A6: 120 Gb/s 12x Small Form-factor Pluggable (CXP)," available for download from the IBTA website.

Table 13-1. 100GBASE-CR10 signals and CXP contact positions

Tx lane	Contact	Tx lane	Contact	Rx lane	Contact	Rx lane	Contact
Signal GND	A1	Signal GND	B1	Signal GND	C1	Signal GND	D1
SL0<pos>	A2	—	B2	DL0<pos>	C2	—	D2
SL0<neg>	A3	—	B3	DL0<neg>	C3	—	D3
Signal GND	A4	Signal GND	B4	Signal GND	C4	Signal GND	D4
SL2<pos>	A5	SL1<pos>	B5	DL2<pos>	C5	DL1<pos>	D5
SL2<neg>	A6	SL1<neg>	B6	DL2<neg>	C6	DL1<neg>	D6
Signal GND	A7	Signal GND	B7	Signal GND	C7	Signal GND	D7
SL4<pos>	A8	SL3<pos>	B8	DL4<pos>	C8	DL3<pos>	D8
SL4<neg>	A9	SL3<neg>	B9	DL4<neg>	C9	DL3<neg>	D9
Signal GND	A10	Signal GND	B10	Signal GND	C10	Signal GND	D10
SL6<pos>	A11	SL5<pos>	B11	DL6<pos>	C11	DL5<pos>	D11
SL6<neg>	A12	SL5<neg>	B12	DL6<neg>	C12	DL5<neg>	D12
Signal GND	A13	Signal GND	B13	Signal GND	C13	Signal GND	D13
SL8<pos>	A14	SL7<pos>	B14	DL8<pos>	C14	DL7<pos>	D14
SL8<neg>	A15	SL7<neg>	B15	DL8<neg>	C15	DL7<neg>	D15
Signal GND	A16	Signal GND	B16	Signal GND	C16	Signal GND	D16
—	A17	SL9<pos>	B17	—	C17	DL9<pos>	D17
—	A18	SL9<neg>	B18	—	C18	DL9<neg>	D18
Signal GND	A19	Signal GND	B19	Signal GND	C19	Signal GND	D19

While there are currently no 100GBASE-CX10 transceivers on the market, there are InfiniBand cables available with attached CXP connectors, which are capable of carrying the signals needed by 100GBASE-CX10. Drawings of the CXP module and mini-multilane connector can also be found in the MSA specifications.

The InfiniBand cable and its attached CXP module provide an example of what a 100GBASE-CX10 transceiver will look like, when a vendor chooses to make one available for short reach connections. Given the cost and size of 100 Gb/s Ethernet ports, it will be some while before providing lower-cost 100 Gb/s Ethernet ports with 100GBASE-CX10 connections becomes economically feasible.

Figure 13-4 shows a cable with CXP modules on each end. The card edge connections inside the CXP module mate with the mini-multilane connector mounted inside the switch port, providing the 84 contact positions, of which the 100GBASE-CR10 standard uses 20 for signals and 14 for signal grounds at each end of the cable.

Figure 13-4. 100GBASE-CR10 CXP connectors and cable

100GBASE-CR10 Signal Encoding

The direct attach cables and the CXP connector module use electrical signaling defined in the standard as a "low-swing AC coupled differential interface." This type of differential signaling provides noise immunity and reduced levels of electromagnetic interference. The differential signal results in a voltage on the wire that is roughly two volts, peak to peak. There are 10 pairs of conductors carrying signals in each direction, for a total of 20 pairs of conductors, or 40 wires, in the cable segment.

The 100GBASE-CR10 link transfers 10 lanes of encoded and scrambled data at 10.3125 Gbaud, to accommodate both data and the overhead associated with 64B/66B coding. The self-clocked nature of the signal eliminates skewing between clock and data signals. The electrical interface to the media is based on a 100 ohm cable impedance, while the signal termination electronics provide both differential and common-mode suppression of signal noise and reflections. The direct attach copper cable is designed to meet a bit error rate specification of 10^{-12}, which is a potential for 1 bit error in every trillion bits sent.

100 Gigabit Ethernet Fiber Optic Media Systems

The 802.3ba supplement provides two 100 Gb/s optical standards. The 100GBASE-SR4 optical system multiplexes the 20 PCS lanes into 10 lanes for transmission over the media, while the 100GBASE-LR4 optical media system multiplexes the PCS lanes into 4 lanes for transmission over 4 wavelengths of light.

The first 100 Gb/s transceivers have been based on the C form-factor pluggable (CFP) module, which is a large module that is capable of handling up to 24 watts of power dissipation. First-generation transceivers with multiple chips and larger power requirements have used this module, which is specified by a multisource agreement.[2]

Figure 13-5 shows a CFP module, which can be used to provide either a 100GBASE-SR10 or a 100GBASE-LR4 transceiver. This figure shows a 100GBASE-LR4 module, which provides two SC fiber optic connectors for connection to a pair of single-mode fibers. The operation of the 100GBASE-LR4 connection is described later in this chapter.

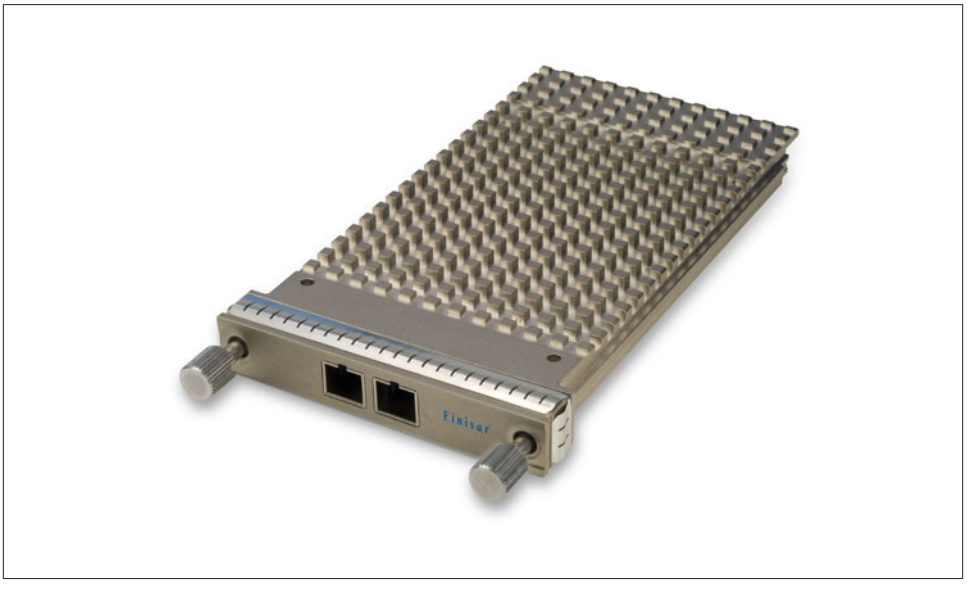


Figure 13-5. 100 Gigabit CFP transceiver module

As the 100 Gb/s media interfaces evolve, they are becoming more efficient and are being upgraded with faster internal paths for transmitting signals within the transceiver and onto printed circuit cards. Recent improvements in electrical signaling standards have provided electrical signals that can achieve 25 Gb/s on high-volume chip and card in-

2. Revision 1.4 of the CFP MSA Hardware Specification can be found on the CFP website.

terfaces. The set of Common Electrical I/O (CEI) specifications that can provide up to 25 Gb/s for electrical signaling were first published by the Optical Internetworking Forum in 2011.[3] Based on these new standards for electrical signaling, fewer lanes are needed, resulting in fewer connections. For that reason, new generations of transceiver electronics are more compact, and new modules have been developed to reduce the size of the transceiver.

The most commonly used transceiver module for 100 Gb/s operation thus far has been the CFP module; it measures 82 mm (3.22 inches) by 14 mm (0.55 inches), making it a large module that takes up a lot of space on the front panel of a switch or router. New CFP specifications have been developed for a CFP2 module (41 mm wide) and for a CFP4 module (21 mm wide). The CFP2 and CFP4 form factors will make it possible to double and quadruple front panel port density, respectively.

The smaller CFP2 and CFP4 modules use fewer electrical connections into the switch, and thus support fewer lanes. Using fewer lanes means that each lane must signal at a higher rate, such as 25 Gb/s. These new transceiver modules are using the latest technology based on the newer OIF signaling standards, combined with advances in circuit integration that require less power for the transceiver. CFP2 modules were first demonstrated in 2012, and CFP4 modules were demonstrated in October 2013. Adoption of these new modules by vendors will occur as they are placed into production and can be purchased in volume quantities, and as vendors integrate these new designs into their equipment.

Cisco CPAK Module for 100 Gigabit Ethernet

One major vendor, Cisco Systems, acquired new module technology with the purchase of a company called Lightwire, that has been developing "silicon photonics" based on complementary metal-oxide semiconductor (CMOS) technology for high-speed networking. This acquisition made it possible for Cisco to rapidly develop its own module, called the CPAK, as an alternative to the CFP2.[4] The CPAK module is only 35 mm (1.37 inches) wide—70% smaller than the original CFP—and also narrower than the new CFP2 module. According to the Cisco data sheet, the CPAK 100GBASE-LR4 module operates at less than 5.5 watts.

Figure 13-6 shows a Cisco 100GBASE-LR4 CPAK module. The smaller size and lower power requirements of this module make it possible for Cisco to provide multiple 100 Gigabit Ethernet ports on the front panel of a fixed switch, or on a switching module in a chassis switch.

3. Optical Internetworking Forum, "Common Electrical I/O (CEI)--Electrical and Jitter Interoperability agreements for 6G+ bps, 11G+ bps and 25G+ bps I/O," September 1, 2011.

4. See the Cisco product sheet on the "CPAK 100GBASE Modules" and the Cisco white paper "Cisco CPAK for 100Gbps Solutions."

Figure 13-6. Cisco CPAK 100 Gigabit transceiver module

100 Gb/s Fiber Optic Media Specifications

The 100 Gigabit standard provides specifications that the multimode and single-mode fiber optic media must meet to support 100 Gigabit Ethernet.

100GBASE-SR10 short-reach media system specifications

The 100GBASE-SR10 short reach system sends 10 lanes of PCS data over 10 pairs of multimode cables, for a total of 20 cables. The 100GBASE-SR10 Ethernet modules provide a 24-fiber multifiber push-on (MPO) jack, to which you connect an MPO plug to provide a connection to the 10 pairs of fiber optic cables. The standard provides three alternatives for MPO plug connectors to make connections over a 100GBASE-SR10 link.

Figure 13-7 shows the three options. The recommended option is to use a 24-fiber MPO connector for all connections in the link. However, for existing cabling systems that are based on 12-fiber MPO connectors, two MPO connectors can be used to provide the total set of 10 lanes and 20 fiber optic cables that is required. The MPO connectors each

have two alignment pins on the plug and two alignment holes on the jack, which help keep the connectors and their fibers correctly aligned when the connectors are mated.

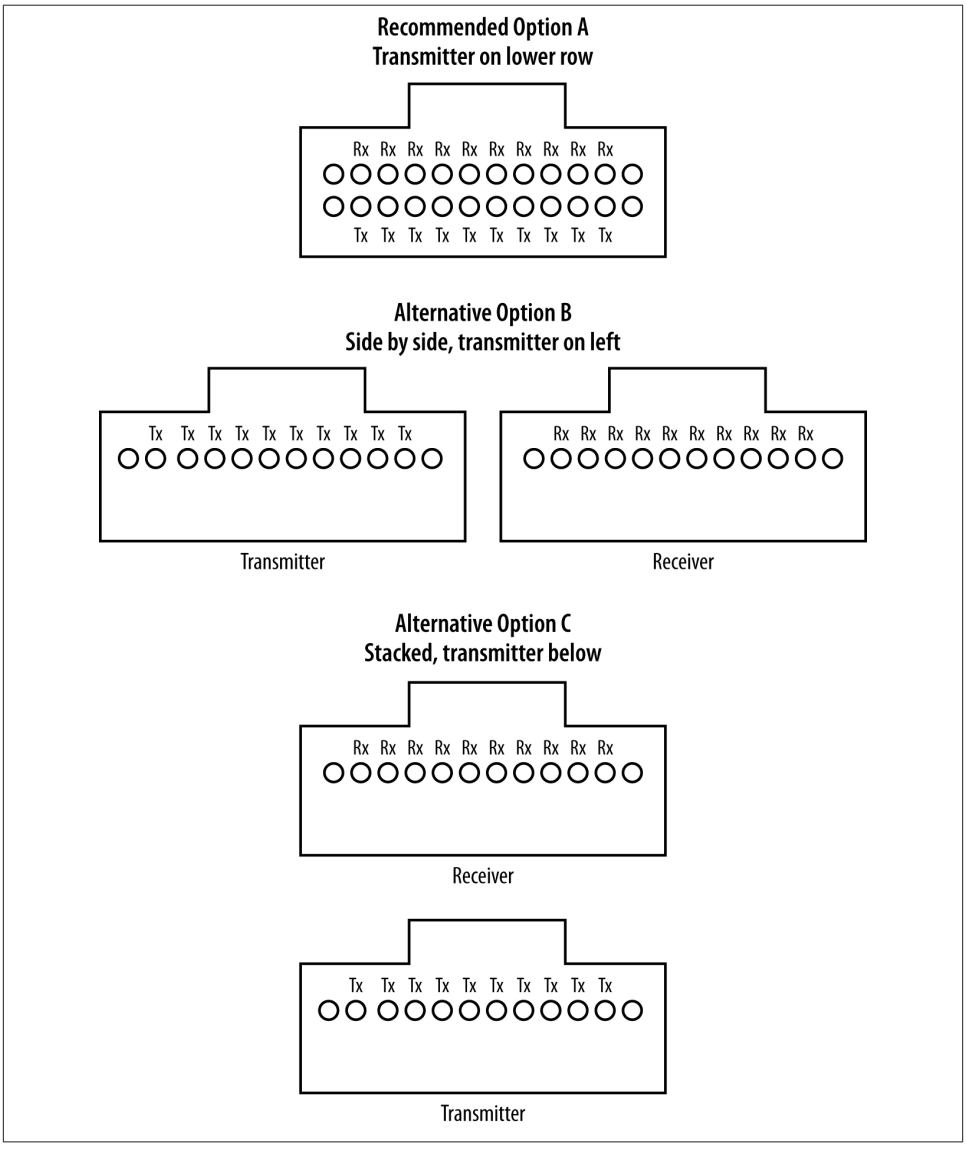


Figure 13-7. MPO connections for 100GBASE-SR10

If you are connecting an optical jumper cable from a CFP module in a switch port directly to a server interface, then you need a short cable with the correct cable connectors on each end. For 100GBASE-SR10, this cable would have 24-fiber MPO con-

nectors on each end. More information on MPO cables and connectors can be found in Chapter 17.

The 100GBASE-SR10 media type is based on multimode fiber optic cables. Multimode fiber optic components are less expensive than single-mode components, and are able to transmit light for relatively short distances compared to single-mode systems.

Table 13-2 shows the distances and channel insertion loss for 100GBASE-SR10. The modal bandwidth of a multimode fiber optic cable refers to its signal-carrying characteristics. Higher modal bandwidth means that the cable can transmit a signal for a longer distance while maintaining sufficient signal quality to the optical receiver in the Ethernet interface at the end of the link. The 50 µm refers to the diameter of the signal-carrying portion of the fiber optic cable.

Table 13-2. Optical specifications for 100GBASE-SR10

Fiber type	Minimum modal bandwidth at 850 nm (MHz-km)	Channel insertion loss (dB)	Operating range (m)
50 µm MMF OM3	2,000	1.9	0.5 to 100
50 µm MMF OM4	4,700	1.5	0.5 to 150

You may wonder why the maximum media lengths for 100GBASE-SR10 are shorter than the lengths specified in the 10GBASE-SR media system (see Table 11-5). Given that the 10GBASE-SR maximum segment lengths are 300 m and 400 m over OM3 and OM4 fiber, respectively, and given that the 100GBASE-SR10 lanes are based on the same technology used in 10GBASE-SR, then why can't the 100GBASE-SR10 system provide the same maximum segment lengths?

The answer is that the transceiver specifications were changed in 100GBASE-SR10 to allow for less-expensive optical components to be used, given that there are 10 transmitters and receivers in each transceiver. The major changes were to relax some of the timing requirements, and to loosen certain optical transmission specifications. The maximum segment length was reduced, in turn, to maintain signal quality when using the lower-cost optical transmitters and receivers.

The total link segment power budget for both OM3 and OM4 fiber types is 8.3 dB, with varying amounts of power consumed by optical noise characteristics, intersymbol interference issues, and the like, depending on the type of the multimode fiber. The channel insertion loss is that portion of the link segment optical power budget that can be consumed by the optical cabling and connectors used on a given segment. As long as the optical power loss, as measured through the link segment from end to end, is at or below the level shown for the channel insertion loss, then the segment will operate correctly.

100GBASE-LR4 long-reach media systems specifications

The 100 Gb/s long-reach media systems use a single pair of fiber, on which four lanes of data are sent over four wavelengths of light.

If you are connecting a CFP module in a switch port to an optical connector termination point in your data center or cabling closet, the optical connectors in your termination point may be different than the ones used on the 100 Gb/s transceiver module. That will require a cable with a different type of optical connector on each end.

Maximum segment lengths are dependent on a number of factors. Fiber optic segment lengths will vary depending on the cable type and wavelength used by the media type. More information on multimode and single-mode fiber optic segments and components can be found in Chapter 17.

The 100GBASE-LR4 media system was designed to operate over single-mode fiber optic cables, with the specifications shown in Table 13-3. The specifications for the long-reach optical fibers used in the 100GBASE-LR4 are simpler, because the transmission characteristics for single-mode fiber optic cables do not include the modal bandwidth considerations found in multimode fiber specifications.

Table 13-3. Optical specifications for 100GBASE-LR4

Optical type	Channel insertion loss (dB)	Operating distance
9 µm SMF	6.3	2 m to 10 km

The total link power budget for the 100GBASE-LR4 system is 8.5 dB, with varying amounts of power consumed by optical noise characteristics, intersymbol interference issues, and the like. The channel insertion loss is that portion of the link segment optical power budget that can be consumed by the optical cabling and connectors used on a given segment.

100GBASE-LR4 wavelengths

The four lanes of PCS data are sent over four wavelengths of light in the 100GBASE-LR4 long reach over single-mode media system. All four wavelengths of light are sent over a single pair of fiber optic cables, using a system called coarse wave division multiplexing (CWDM). Each wavelength of light has a specific CWDM frequency, as specified in the wavelength grid defined in the ITU-T G.694.2 standard.

Table 13-4 shows the four center wavelengths, and the frequency range for each of the four wavelengths, or "colors," of light used to carry signals over the single-mode cable. Each 100GBASE-SR4 transceiver contains a four-wavelength optical transmitter and a four-wavelength receiver. The single-mode optical cable simply carries the four wavelengths of light between the transceivers, providing four paths for the four lanes of PCS data, with the 64B/66B encoding resulting in a signal rate of 25.78125 Gbaud per second.

As we've described, the four PCS lanes carry the Ethernet frame data between the interfaces so that the resulting data rate for an Ethernet frame transmission is 100 Gb/s.

Table 13-4. 100GBASE-LR4 wavelengths

Lane	Center wavelength	Wavelength range
L_0_	1,295.56 nm	1,294.53 to 1,296.59 nm
L_1_	1,300.05 nm	1,299.02 to 1,301.09 nm
L_2_	1,304.58 nm	1,303.54 to 1,305.63 nm
L_3_	1,309.14 nm	1,308.14 to 1,310.19 nm

100GBASE-ER4 media specifications

The 100GBASE-ER4 media system was designed to operate over single-mode fiber optic cables, with the specifications shown in Table 13-5.

Table 13-5. Optical specifications for 100GBASE-ER4

Optical type	Channel insertion loss (dB)	Operating distance
9 µm SMF	15	2 m to 30 km
9 µm SMF	18	2 m to 40 km [a]

[a] Links longer than 30 km are considered "engineered links." Fiber cable attenuation for such links must be less than 0.43 to 0.5 dB/km. Using fiber with 0.5 dB/km attenuation may not support operation at 10 km for 100GBASE-LR4 or 40 km for 100GBASE-ER4.

The total link power budget for the 100GBASE-ER4 system is 21.5 dB, with varying amounts of power consumed by optical noise characteristics, intersymbol interference issues, and the like. The channel insertion loss is that portion of the link segment optical power budget that can be consumed by the optical cabling and connectors used on a given segment.

CHAPTER 14
400 Gigabit Ethernet

The development of a higher-speed Ethernet system beyond 100 Gb/s began with an IEEE "Bandwidth Assessment Ad Hoc group," which was announced in May 2011. This group held multiple meetings and teleconferences over the period of a year, during which they analyzed the bandwidth requirements for a range of customers and industries.[1] The group found that bandwidth requirements were growing at an average of 58% annually, and that there was an immediate need to begin development of a higher-speed Ethernet system. These findings were documented in a report published in July 2012.[2]

Following the publication of the Bandwidth Assessment report, a "Higher Speed Ethernet Consensus Ad Hoc" group was formed to develop a consensus on what speed could be achieved in the next couple of years for a higher-speed Ethernet system.[3] The consensus group concluded that 400 Gb/s was technically achievable, and that waiting for 1 terabit (Tb) technology to be developed would delay the creation of a new standard by several years. Experts in optical components and other signaling elements reported that achieving rates beyond 400 Gb/s was not possible with current production-quality components. They determined that to achieve 1 Tb/s speeds outside the laboratory would require major investments in research and development. And that, in turn, meant that waiting for 1 Tb/s would result in a considerably longer time to market.

1. The meeting reports can be found in the public area of the IEEE website.
2. The report is called "IEEE 802.3 Industry Connections Ethernet Bandwidth Assessment".
3. Refer to the meeting notes for the Higher Speed Ethernet Consensus Ad Hoc group.

400 Gb/s Ethernet Study Group

The conclusion to proceed with the development of a 400 Gb/s speed for the next generation of Ethernet led to the formation of yet another group, this one chartered to dive into the technical details and confirm that 400 Gb/s was an achievable goal for a new IEEE Ethernet standard. To that end, the creation of a 400 Gb/s Ethernet study group was announced in April 2013.

The 400 Gb/s study group is proceeding to develop the technical information that will lead to a formal Project Authorization Request (PAR). When a PAR is granted, that launches the formal beginning of a new standardization effort. At that point, the 400 Gb/s Ethernet specifications will be assigned an IEEE 802.3 supplement number, which is expected to be 802.3bs, and a schedule will be developed for the completion of the new standard. Predicting when a new 400 Gb/s standard will be completed is difficult, given the amount of work that needs to be done. The study group's chairman, John D'Ambrosia, believes that it is possible that the new standard may be formally adopted sometime in 2017.[4]

Assuming that good progress is made on the new standard, it's not unlikely that the specifications could be substantially complete in 2016.

400 Gb/s Standardization

While this process may seem to include creating a lot of groups and holding a lot of meetings, it helps to remember that the IEEE standards are developed by consensus among a wide range of stakeholders. These stakeholders include the component manufacturers, who have to build the optical and electronic components to operate at these new speeds; the vendors, who must be able to build economically viable switches, routers, and other Ethernet devices that will sell at prices customers can afford; and the customers themselves, who intend to use Ethernet in a wide variety of environments and who expect Ethernet technology to be fast, reliable, and easy to use.

There are a lot of elements involved in successfully developing a new standard for the marketplace, and a considerable amount of time and effort goes into each new standard, not to mention literally thousands of hours spent by hundreds of engineers and others at IEEE meetings and in development work.

Proposed 400 Gb/s Operation

The 400 Gb/s Ethernet study group is considering various ways in which the new higher-speed Ethernet system could work. These include faster transmission schemes, more

4. Stephen Lawson, "Ethernet's 400-Gigabit challenge is a good problem to have," *ComputerWorld*, October 15, 2013.

complex modulation mechanisms, and transmitting more lanes of data over the media system.

Transmitting 16 lanes of data at 25 Gb/s is one way to achieve a 400 Gb/s Ethernet link, while still using technology that is well known and that is available at reasonable cost. It's not unlikely that the first generation of a 400 Gb/s optical interface standard could be based on sending 16 lanes of data. As technology evolves and the new 50 Gb/s signaling standard becomes more widely available, it's possible that a second generation 400 Gb/s standard using eight lanes operating at 50 Gb/s could also be developed. Other mechanisms for sending data at 400 Gb/s are also under active consideration, and it is too early to tell which methods will be adopted in the final specifications.

PART III
Building an Ethernet System

Part III will describe how to build Ethernet local area networks. Chapter 15 covers the structured cabling standards and how structured cabling systems are organized. Chapters 16 and 17 explain how twisted-pair and fiber optic cables and connectors work, and how to use them.

CHAPTER 15
Structured Cabling

An essential truth of networking is that an Ethernet system can never be better than its cabling. Providing high-quality cabling can be easy enough for a small system based on a single Ethernet switch in a home that supports a small number of devices. You can connect the devices to the switch ports with high-quality patch cords, and your network will be complete. However, the majority of networks support more than just a few devices. Instead, office buildings these days require network systems that connect to practically every room in the building. Providing a high-quality cabling system for an entire building is a much more complex task. That's where a structured cabling system can help.

Structured cabling systems achieve their goals by providing a cabling hierarchy based on backbone cables that carry signals between telecommunications closets on various floors of a building, and horizontal cables that deliver data from the telecommunications closet to the networked device. The ease with which such a system can be expanded and rearranged helps manage the moves, adds, and changes that occur.

A structured cabling system is based on point-to-point cable segments that are installed according to detailed guidelines and specifications published in the structured cabling standards. This provides a reliable and manageable cabling system. A structured cabling system based on industry standards and high-quality components allows your network to function at its best, delivering stable network services for your users day in, day out.

You can think of the cabling system as the essential skeleton of your network. Like most skeletons, it's typically hidden out of sight, which means that it can easily be forgotten. Overlooking your cabling system can be dangerous, however, as the lack of a solid and well-designed cabling system can lead to unstable network operation and make network growth and management much harder to accomplish.

Despite the importance of building a high-quality media system, you won't find any advice in the Ethernet standard on how to proceed with this task. That's because the

specifications for structured cabling systems are outside the scope of the Ethernet standard. However, cabling systems are very much inside the range of things a network designer must accomplish in order to build a reliable and manageable network system.

This chapter provides an introduction to and overview of the structured cabling standards, and shows how Ethernet cabling fits within those standards. The basic elements of a structured cabling system are described, with emphasis on the horizontal cabling segments used to connect Ethernet stations to switches. We also describe the new cable specifications and testing standards that have been developed to support high-speed Ethernet systems, including 1 and 10 Gigabit Ethernet systems.

Note that this chapter can only provide a brief overview of a very large topic. There are many standards involved when cabling a building, including not only structured cabling standards but also electrical safety standards, fire safety standards, zoning regulations, and more. A full treatment of structured cabling practices, rules, and regulations would occupy an entire bookshelf. As discussed later in this chapter, it is strongly recommended that large cabling projects be installed by trained professionals who have studied structured cabling, are certified in the standards and practices, and have been trained to do the job correctly.

Structured Cabling Systems

A major advantage of a structured cabling system is that it is designed to make it easier to deal with the constant task of handling moves, adds, and changes. A structured cabling system is also designed to provide a flexible cabling system that can support devices of all kinds, including desktop computers, network-based Voice over IP (VoIP) telephones, and wireless access points. Finally, a structured cabling system results in a more reliable network that is also easier to troubleshoot when a failure occurs.

A set of cables installed without any particular plan or regard for the industry guidelines might appear to work well enough at first. However, the lack of structure will make it difficult to accommodate network growth and to troubleshoot the system when problems arise.

When designing a cabling system, it's important to plan. The goal is to come up with a plan that scales well, and that will accommodate constant growth while still maintaining order. You also want to make sure that the cabling used in your system can accommodate higher network speeds as required. Network systems are almost always growing and changing to accommodate new technology, add new connections, and allow people to move around. These tasks are collectively referred to as "moves, adds, and changes" (MAC), and you may hear references to the "MAC cycle" by facilities managers.

Because this is easily confused with the media access control (MAC) layer of the Ethernet standards, this particular piece of jargon is best left to facilities managers.

Unstructured cabling systems built without reference to industry standards are often prone to intermittent network failures that can come and go depending on the time of day and the traffic load. Troubleshooting a cable system that "just grew" without any particular structure can be a time-consuming process of tracing cables to their source. Meanwhile, the users are unproductively waiting for the network to come back up. A structured cabling system can help avoid these problems.

The ANSI/TIA/EIA Cabling Standards

In this chapter, we will focus on the ANSI/TIA/EIA standards that are widely used in the United States. The ANSI/TIA/EIA standards are a set of vendor-independent structured cabling specifications that have been developed by two trade organizations, beginning with the Electronic Industries Association (EIA) and continuing with the Telecommunications Industry Association (TIA). Both the TIA and the EIA are members of the American National Standards Institute (ANSI), which is the coordinating body for voluntary standards groups within the United States. These specifications are periodically adopted by the American National Standards Institute, to become the ANSI/TIA/EIA Telecommunications Standards. Websites providing copies of these standards for sale are listed in Appendix A.

The most recent version of the structured cabling specifications is officially called the ANSI/TIA-568 family of telecommunications standards. We'll look at some of these shortly.

The goal of the TIA cabling standards is to provide a vendor-independent cabling system supporting both voice and data requirements. Before the creation of the TIA standards, there was no nonproprietary standard you could turn to for guidance when it came to installing a cabling system in your building.

Solving the Problems of Proprietary Cabling Systems

When Ethernet was first developed in the 1980s, building cabling systems were primarily designed to support telephones and included only voice-grade twisted-pair cable suitable for telephone communications. If you wanted to support data communications, you typically had to install a proprietary cabling system from a computer vendor. If you needed to support multiple network systems, then your building ended up with a ceiling stuffed full of bulky cables installed to support equipment from several different computer vendors. Each vendor's cable system used noncompatible cables and connectors.

In those days, it seemed that every vendor had a different approach to cabling for its computer network equipment. Facilities managers were forced to invent their own procedures and policies for dealing with the unholy tangle of cables that could result.

The TIA standards helped solve this problem by providing a single source of specifications and recommendations for a set of structured cables capable of handling everything your building might need to support. For example, the ANSI/TIA-568-C.0 cabling standard for commercial buildings specifies component requirements, cabling distances, and outlet and connector configurations. The standard also provides recommended cabling topologies. By using these standards, you can design structured cabling systems to support all manner of telecommunications, including voice, data, and video.

The complete set of standards for structured cabling systems is comprised of a series of documents, and as with all standards, there is a constant process of revision and change. The horizontal cabling specifications provided in ANSI/TIA-568-C.0 are the portion of the standard that you will encounter most often when connecting Ethernet equipment. Therefore, we will provide a detailed description of the horizontal segment later in this chapter.

ISO and TIA Standards

You should know that there is also an international cabling standard developed by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), called ISO/IEC 11801, Edition 2.2, "Information technology - Generic cabling for customer premises." This standard covers the same general range of topics as the ANSI/TIA-568 set of standards, and includes its own rating system for cables. The ISO standard qualifies the performance of network channels and links with the term *class*, and lists classes of performance including C, D, E, E_A, F, and F_A. The 11801 standard also specifies optical cable channel and link classes of OF-300, OF-500, and OF-2000.

While the technical specifications provided in the TIA and ISO standards are similar for various grades of cabling, the terminology used in the standards varies, which can cause confusion. In the TIA standards, the cabling components are characterized in terms of a performance "category," and the links or channels built with those components are also described with a performance category. In the ISO standards, while cables and components are characterized by performance categories, the links and channels built with those components are characterized by a performance "class."

The ANSI/TIA Structured Cabling Documents

A few years ago, the TIA standards were extensively revised and reorganized. The reorganization led to a set of new documents that were published beginning in 2009. These include:

ANSI/TIA-568-C.0 "Generic Telecommunications Cabling for Customer Premises"
: This standard describes the planning and installation of cabling for all types of customer premises, by specifying a generic cabling system. This includes cabling system structures, basic topologies and distances, installation, performance and testing, and optical fiber transmission and test requirements.

ANSI/TIA-568-C.1 "Commercial Building Telecommunications Cabling Standard"
: This standard describes more specific details for planning and installing a structured cabling system within a commercial building, and between commercial buildings in a campus environment.

ANSI/TIA-568-C.2 "Balanced Twisted-Pair Telecommunication Cabling and Components Standard"
: This standard includes the component and cabling specifications and the testing requirements for copper cabling (including Category 3, 5e, 6, and 6A cables). The standard recommends Category 5e cabling to support 100 MHz operations, which covers all Ethernet speeds up to 1 Gigabit.

ANSI/TIA-568-C.3 "Optical Fiber Cabling Components Standard"
: This standard includes the cable and component specifications for premises optical fiber. Although the standard is primarily intended for use by manufacturers, others (such as cabling system designers, installers, and end users) may find it useful.

Elements of the Structured Cabling Standards

The 568-C.1 commercial building cabling standard enumerates several basic elements of a structured cabling system. We'll quickly list these elements, because you will often encounter these terms when dealing with cabling systems. Following this, we show how these basic elements are used in a star topology, which is the basis of the structured cabling standards. Terms to be familiar with include:

Building entrance facilities
: The cables, surge protection equipment, and connecting hardware that may be used to link the cabling inside your building with the campus data network are located here.

Equipment room
: This is a space reserved for more complex equipment than may be found in telecommunications closets. Equipment rooms may be used for major cable terminations and for any grounding equipment needed to make a connection to the campus data network and public telephone network.

Building backbone cabling
: Building backbone cabling based on a star topology is used to provide connections between telecommunications closets, equipment rooms, and the entrance facilities.

Telecommunications room
: The primary function of the telecommunications room, also called a telco closet or wiring closet, is to provide a location for the termination of the horizontal cable on a given floor of a building. This closet houses the mechanical cable terminations and any cross-connects for the horizontal and backbone cabling system. It may also house interconnection equipment, including Ethernet switches.

Horizontal cabling
: The horizontal cabling system extends from the telecommunications room to the telecommunications outlet located in the work area. Horizontal cabling components include the work area outlets, the horizontal link cables installed between the telecommunications room and the outlets, and cable termination equipment such as patch panels, located in the telecommunications room. Also included are any patch cables required for cross-connects between Ethernet switches in the telco closet and the horizontal cabling system.

Work area
: The work area may be an office space or any other area where computers and other equipment are located. The work area components of a structured cabling system include any patch cables required to connect a user's computer, telephone, or other device to the communications outlet on the wall.

Multiuser Telecommunications Outlet Assembly (MUTOA)
: An optional component for open office environments, the MUTOA provides a termination point for cabling in an open area to allow the cables to then be routed through pathways that are provided in modular office walls.

Star Topology

The structured cabling system described in the cabling standards is based on a star topology. A star topology is a set of point-to-point links originating from a central hub; the links appear to radiate out from the hub like rays from a star.

Figure 15-1 illustrates the basic elements of a structured cabling system, and shows how these elements are arranged in a star topology. The dotted lines indicate equipment rooms, telecommunications closets, and work areas. The cabling standards specify a backbone system with a star cabling topology that has no more than two levels of hierarchy within a building. This means that a cable should not go through more than one intermediate cross-connect device between the *main cross-connect* (MC) located in an equipment room and the *horizontal cross-connect* (HC) located in a wiring closet.

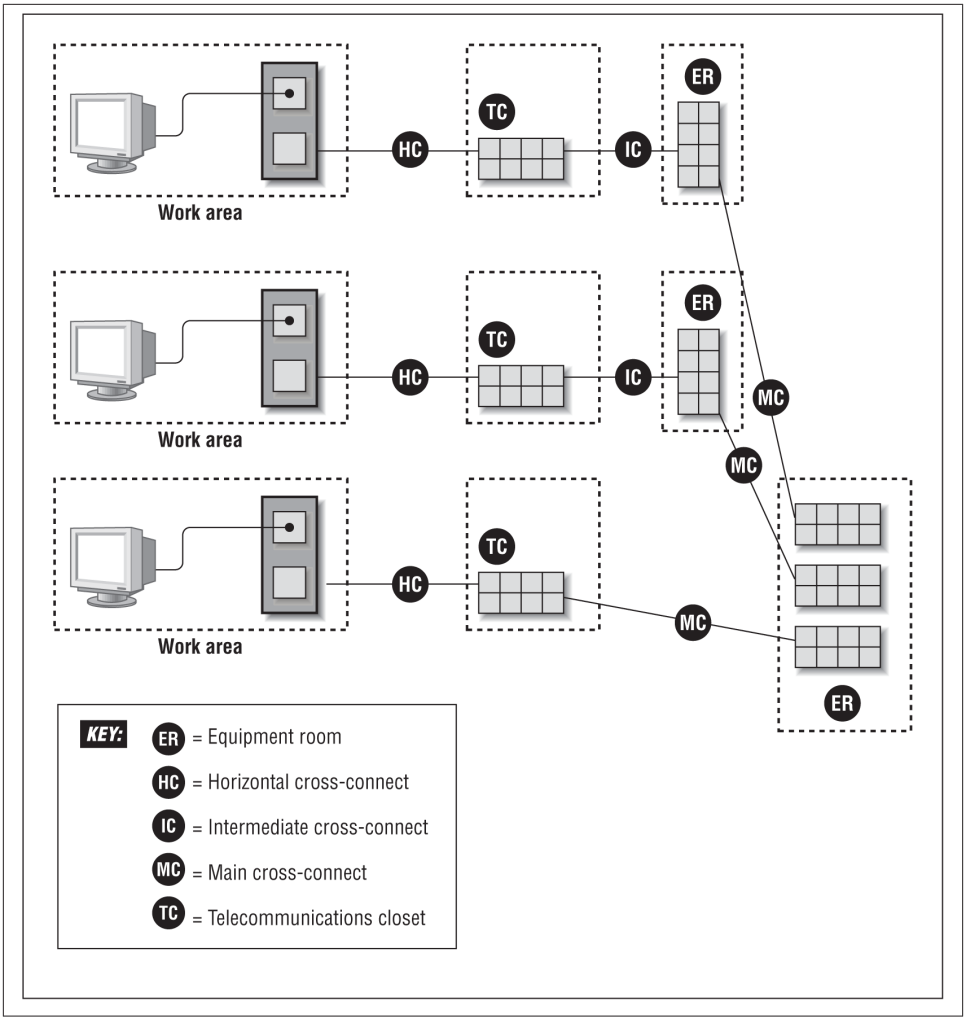


Figure 15-1. Elements of a structured cabling system

There are many advantages to a star topology:

- Central wiring locations ease the task of managing moves, adds, and changes.
- A small number of central cabling points enables faster troubleshooting. If you're holding on to one end of a network connection in an office area, you can know exactly where the other end is located, because all connections in a given area will terminate in the same telco closet. That also means that anyone working on the network can easily determine what equipment is connected to each end.

- Independent point-to-point links prevent cable problems on any given link from affecting other links.
- Central equipment locations provide for easier migration to new technologies. You can, for example, increase the speed of a network by upgrading equipment in a few locations without having to recable the entire building.
- Physical security (e.g., door locks) can be provided for critical equipment that could cause widespread network failure if tampered with.

As you can see, a star topology is a major advantage for an installer or troubleshooter. Star topologies make installing, testing, and troubleshooting network segments much easier and less time-consuming.

Twisted-Pair Categories

The ANSI/TIA-568 family of standards rate twisted-pair cable in terms of which *category* of cable specifications it meets. The category specifications in this set of standards are widely used in the network industry to identify which types of twisted-pair cable can accommodate the various Ethernet media system speeds. There have been a number of cable categories over the years, including:

Category 1 and 2
: Category 1 and 2 cables and connecting hardware are not recognized as part of the cabling standards. These two categories are for older cables used in telephone systems; they are not recommended for carrying Ethernet signals.

Category 3
: This designation applies to 100 ohm unshielded twisted-pair (UTP) cables and associated connecting hardware whose transmission characteristics are specified up to 16 MHz. Category 3 UTP cables with a 100 ohm impedance rating are capable of supporting 10BASE-T Ethernet media systems.

Category 4 and 5
: The Category 4 designation applies to 100 ohm UTP cables and associated connecting hardware with transmission characteristics specified up to 20 MHz, while Category 5 was specified up to 100 MHz. Category 4 was originally designed to support 16-Mb/s token ring systems, and is no longer widely used. Category 5 has been superseded by Category 5e and is no longer recommended by the TIA cabling standards for new installations. However, a number of installed cabling systems are based on Category 5 cabling; these systems function perfectly well for Ethernet speeds up to and including 1000BASE-T. Information on the qualification of legacy Category 5 installations can be found in Annex M of ANSI/TIA-568-C.2.

Category 5e
: Category 5e cabling supersedes Category 5, and was specified in 2000 to include improved technical specifications for the support of 1000BASE-T Ethernet media systems. The 5e specifications increased performance limits for near-end crosstalk (NEXT), equal level far-end crosstalk (ELFEXT), and return loss, among other items.

Category 6
: Category 6 cabling is specified to handle signaling rates up to 200 MHz. The original development aim was to create a higher-quality cable that could "future-proof" a cabling system, as a large increase in twisted-pair speed above 200 MHz was not expected. Unfortunately for this approach, even higher speeds have since been reached, which has made it necessary to develop even higher performance versions of twisted-pair cabling.

Category 6A
: Category 6A cabling was developed to address the higher signaling rates and increased alien crosstalk specifications required to support 10GBASE-T for up to 100 meters of horizontal cabling containing up to four connectors. Category 6A is specified for rates up to 500 MHz and is recommended in the TIA standards as the minimum grade of cabling for 10GBASE-T media systems.

Category 7 and 7A
: Category 7 and 7A cables are specified in the ISO cabling standards. These cables are higher performance, but the higher performance levels are not required for 10GBASE-T operation and are not sufficient for 40GBASE-T operation.

Table 15-1 provides a comparison between the TIA and ISO cabling categories, and the ISO class system that is used to categorize the resulting channel or link segment. Note that Category 6A is sufficient to carry 10 Gb/s Ethernet signals, and there is no requirement for higher-performance cabling to carry 10GBASE-T signals. There's nothing to prevent you from using higher-performance cabling for 10GBASE-T, but it's not necessary.

Table 15-1. Comparison of TIA and ISO copper cable specifications

Max bandwidth	TIA (cabling/components)	TIA (channel/link)	ISO (cabling/components)	ISO (channel/link)
100 MHz	Category 5e	Category 5e	Category 5e	Class D
250 MHz	Category 6	Category 6	Category 6	Class E
500 MHz	Category 6A	Category 6A	Category 6A	Class E_A
600 MHz	N/a	N/a	Category 7	Class F
1,000 MHz	N/a	N/a	Category 7A	Class F_A

The TIA did not specify any category of cable beyond 6A until recently; in 2013, work began on a next generation of cable specifications that will meet the requirements for

carrying 40GBASE-T signals. When the specifications are completed, there will be a new TIA Category 8 twisted-pair cable type suitable for 40GBASE-T systems.

Minimum Cabling Recommendation

The current set of TIA 568 cabling standards recognize that different environments may require different cabling. A data center that is intended to support high-performance servers will require Category 6A cabling to connect to those servers. A typical office building, on the other hand, may be well-enough served with Category 5e cabling, providing up to 1000BASE-T service for stations and other devices. It's up to you to decide what your needs are, and how much you should invest in cabling to meet those needs.

Another consideration is the lifecycle of the cabling plant. If you are designing a new cabling system with the goal of achieving the maximum 10-year lifecycle that the cabling standard can provide, then you can more easily justify the cost of using the best cabling, and install Category 6A cables everywhere. A further consideration is that the installation costs are much larger than the material costs, which means that the additional cost of using Category 6A cables will be a small fraction of the total project cost.

Given the rate at which wireless access points have been increasing in speed, you may also want to consider installing Category 6A cables for wireless access point (AP) connections. By providing the APs with Category 6A cables, you can ensure that the network connection will be able to handle all of the expected current and future speeds as new generations of APs replace the current generation.

Ethernet and the Category System

The specifications for twisted-pair Ethernet media systems were written with the category system in mind. When the 10BASE-T twisted-pair Ethernet standard was developed, it was designed to work over lower-quality Category 3 "voice-grade" cables. But as we've just seen, with the development of higher-speed networks, the structured cabling standard was extended to include specifications for Category 5, 5e, and now 6A cables. All Ethernet twisted-pair media systems up to 1000BASE-T can be supported over Category 5 and Category 5e twisted-pair cables and connecting equipment.

The following is a list of the commonly available twisted-pair Ethernet standards and the cable categories that they are specified to work with:

- The 10BASE-T system is specified for two pairs of Category 3 or better cables, including Category 3 or better connecting hardware, patch cables, and jumpers. A Category 3 25-pair cable may be used in the segment as long as the signal specifications for multiple disturber crosstalk are met. Multiple disturber crosstalk is described Chapter 16.

- The 100BASE-TX system requires two pairs of cables that meet the Category 5 specifications (or better). The connecting hardware, patch cables, and jumpers must also meet these specifications.
- The 1000BASE-T system requires four pairs of Category 5, 5e, or better cabling and hardware.
- The 10GBASE-T system requires four pairs of Category 6A cabling and hardware for 100 m segment lengths. Shorter distances are possible over Category 6 cables, up to a maximum segment length of 37 to 55 m, depending on the signaling capability of the horizontal link.

Horizontal Cabling

A horizontal cable is one that extends from the communications outlet in the office or work area to the telecommunications closet. The list of components that may be found in a standard horizontal cabling system include:

Horizontal link cabling

A horizontal link, described in detail later in this chapter, may extend a maximum distance of 90 meters (295 feet). There are two types of cables recognized in the TIA cabling standards as options for use in horizontal links:[1]

- Four-pair (eight wires), 100 ohm impedance twisted-pair cabling. The recommended connector type is an eight-position RJ45-style modular jack terminating all eight wires of the four-pair cable.
- Multimode optical fiber cabling, either 62.5/125 micron (μm) or 50/125 μm. The 50/125 μm 850 nm laser-optimized multimode fiber is recommended. Currently, the most popular recommended connector type is the small form-factor (SFF) LC connector. Another recommended fiber optic connector type is the "SC" connector, formally called an SCFOC/2.5 duplex connector. SC stands for "Subscriber Connector," and FOC stands for "Fiber Optic Connector."

Telecommunications outlet/connector

A minimum of two *work area outlets* (WAOs) is specified for each work area; each work area is connected directly to a telecommunications closet. One outlet should connect to one four-pair (eight-wire) UTP cable. The other outlet may connect to either another four-pair UTP cable or a fiber optic cable, as required to meet the needs of the work area. Any active or passive adapters needed at the work area should be external to the outlet.

1. The 50 ohm coaxial cables used for the original thick and thin Ethernet systems and the shielded twisted-pair Token Ring cables are no longer recognized cable types in the cabling standards.

Cross-connect patch cables

Equipment cable and patch cables used in telecommunications closets should not exceed 6 meters (19.6 feet) in length. An allowance of 3 meters (9.8 feet) is provided for the patch cable from the telecommunications outlet to the workstation. A total allowance of 10 meters (32.8 feet) is provided for all patch cables and equipment cables in the entire length from the closet to the workstation. This, combined with the maximum of 90 meters of horizontal link cable distance, makes a total of 100 meters (328 feet) for the maximum horizontal channel distance from the network equipment in the closet to the computer in the office.

Horizontal Channel and Basic Link

The TIA standards define a basic link and a channel for installation and testing purposes. Figure 15-2 shows a basic link, which consists of the fixed cable that travels between the telecommunications outlet in the office or work area and the wire termination point in the wiring closet. The basic link is limited to a maximum length of 90 meters (295 feet). A cabling contractor can install this portion of the cabling system in a building and test it according to the specifications. This provides a way to certify the installed cabling in a new cabling system before any network equipment is hooked up.

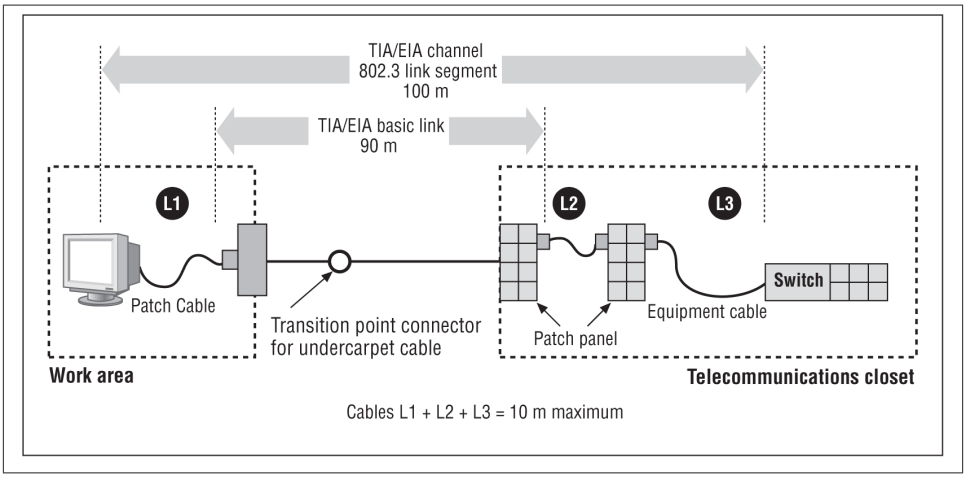


Figure 15-2. Basic link and channel

The standard includes an allowance for an additional maximum length of 10 meters (32.8 feet) for patch cables and equipment cables. These cables can be located between the wiring hub and the patch panel or cross-connect located in the closet. They can also be located between the wall outlet and the computer located in the office or work area.

The total horizontal cable segment including all patch cables and equipment cables is called the *channel*, and may be a maximum of 100 meters (328 feet) in length. The standards include test specifications for the complete channel to allow end-to-end testing that includes all patch cords and equipment cords. The link segment specifications in the IEEE 802.3 standard for UTP Ethernet segments are based on the entire end-to-end channel as well. For example, the maximum amount of signal attenuation specified for a 10BASE-T segment is 11.5 dB, which is the amount of signal loss allowed for the entire horizontal segment from one end of the channel to the other.

The 100 Meter Design Goal

Both the ANSI/TIA/EIA standards and the twisted-pair Ethernet specifications are written with a 100 meter length design goal for segments. The 100 meter goal comes from studies that show that the vast majority of the desktops in the average office building are within 100 meters of cable length from the nearest telecommunications closet.

The cabling standards note that horizontal cabling systems contain the greatest number of individual cables in a building, and that horizontal cabling is typically much more difficult to replace than backbone cabling. That's because horizontal cabling is pulled through the ceilings and walls of the building to reach each work area, making the installation costs of horizontal cabling much higher. For this reason, high-quality cabling and components should be used in the horizontal cabling system.

Cabling and Component Specifications

The horizontal link cable is just one element of a horizontal cabling system. You also need to make sure that all of the components used are rated to meet the correct category specifications. The horizontal channel will typically include a jumper or patch cable from the Ethernet switch in the telecommunications closet to a wire termination patch panel of some sort. From that patch panel, the horizontal link cable travels over ceilings and through the walls until it reaches the work area, typically an office or cube space.

In the work area, the horizontal link cable terminates in another eight-position jack connector installed in a wall plate. To make a connection to a desktop computer or other device, you need to connect another patch cable from the modular jack connector in the wall plate to the modular jack connector in the network interface of the computer.

Every one of these components must be rated for the correct category of signal performance in order to end up with a complete horizontal channel rated at the category you have specified for the building cabling. If all of the cables and components in a segment are not at the correct level of quality, then you may experience slow network performance caused by increased frame loss due to signal errors.

Category 5 and 5e Cable Testing and Mitigation

The 1000BASE-T system is more sensitive to the amount of signal reflections caused by cable connectors, as well as to the amount of signal crosstalk on the cable. This means that existing cable plants should be checked to ensure that they meet the 1000BASE-T requirements.

Existing Category 5 cabling plants that met the Category 5 cabling standards when they were installed and that are currently supporting Fast Ethernet systems should be able to carry 1000BASE-T signals without difficulty. However, you should retest at least a sample of the installed cable segments to make sure that they meet all signal parameters. The TIA cabling standards include testing specifications to characterize the correct operation of Category 5 cabling systems. Testing can be done with handheld cable testers that automatically cycle through the series of tests required to certify links for the performance standards in the specifications.

Some small percentage of Category 5 cabling systems may be found to have been installed incorrectly or with substandard components. These systems are unlikely to be able to support either Fast Ethernet or Gigabit Ethernet signals. If testing uncovers a link that does not meet the specifications, there are several things you can do to take corrective action. The link should be retested after each corrective action is applied:

- Replace the patch cord at the work area end of the link with a patch cord that meets or exceeds the Category 5e specification.
- Reconfigure the link to eliminate wiring closet cross-connect cables and connectors.
- Replace any transition point connector with equipment that meets or exceeds the Catgeory 5e specification.
- Replace the work area outlet with an outlet that meets or exceeds the Category 5e specification.
- Replace the wiring closet interconnect equipment (patch panel) with equipment that meets or exceeds the Category 5e specification.

Cable Administration

The ANSI/TIA-606-B standards (available for purchase via the TIA website) provide a set of specifications for administering cabling systems, including labels and records. A cable identification scheme is essential in any network that contains more than a few stations. A complete floor or building network can quickly get out of hand without a coherent cable labeling scheme. There's no particular magic to coming up with a cable labeling system; you simply need to understand the naming and labeling recommendations, come up with a system that works for you, and follow it consistently.

Getting the labels onto the cables as they are installed can be a major challenge, as retrofitting labels is a very time-consuming and error-prone process. If the cables aren't labeled when they are installed, you will find that it's almost impossible to get around to labeling them later. The memory of where the cables go fades more quickly than you might imagine, and before you know it, you're left with a network system that consists of a maze of twisty little cables that all look alike.

Another difficulty can be finding labels that will stick to the cables and not fall off over time. For that reason, it's best to use labels that are specifically designed for cable systems.

Identifying Cables and Components

The recommended schemes for identifying cables are based on the major components of a cabling system. Because the wire distribution equipment is installed in *equipment frames*, also known as *telco* or *equipment racks*, the cable termination points are called *distribution frames*. The *horizontal distribution frame* (HDF) may consist of one or more telecommunications closets on a floor. The *main distribution frame* (MDF) is the main equipment room where the backbone cabling of the building is terminated.

The basic cable identifications based on the TIA-606-B standard are designed to provide as much information as possible on the cabling system itself, in an effort to minimize the need to keep and refer to external documentation. Your facilities manager may prefer another approach; there is no single system that meets all local requirements.

Class 1 Labeling Scheme

The TIA-606-B labeling scheme is based on the basic elements of a structured cabling system. The following labeling scheme is defined in the standard for a "Class 1 Administration" space, which is the simplest case. A Class 1 space is served by a single equipment room, which is also the only telecommunications space. The standard also provides labeling schemes for more complex spaces with multiple floors and telecommunications spaces. For Class 1 spaces, it includes the following identifiers:

Horizontal link identifier
: A horizontal link identifier for work area outlets is assigned to each horizontal link. The label appears on the cable connector or outlet faceplate where the link terminates. The format is *fs-an*, where:

- *f* = One or more numeric characters identifying the floor of the building occupied by the telecommunications space (TS)
- *s* = One or more alphabetic characters uniquely identifying the TS on floor *f*, or the building area in which the TS is located
- *a* = One or two alphabetic characters uniquely identifying a single patch panel or group of panels with sequentially numbered ports, or an insulation dis-

placement connector (IDC), or a group of IDCs serving as part of the horizontal cross-connect

- *n* = Two to four numeric characters designating the port on a patch panel in the TS or the section of an IDC on which a four-pair horizontal cable is terminated in the TS

Backbone cable identifier
: A unique building backbone cable indentifier is assigned to each backbone cable between two telecommunications spaces in a building. It takes the format of fs_1/fs_2-*n*, where:

  - fs_1 = The TS identifier for the space containing the termination of one end of the backbone cable
  - fs_2 = The TS identifier for the space containing the termination of the other end of the backbone cable
  - *n* = One or two alphanumeric characters identifying a single cable with one end terminated in the TS designated fs_1, and the other end terminated in the TS designated fs_2
  - The TS with the lesser alphanumeric identifier must be listed first. If the entire cable is within one TS, the format may be fs_1/fs_1-*n*

Telecommunications space identifier
: A TS identifier, unique within the building, is assigned to each TS. It has the format *fs*, where:

  - *f* = One or more numeric characters identifying the floor of the building occupied by the TS
  - *s* = One or more alphabetic characters uniquely identifying the TS on floor *f*, or the building area in which the space is located

With this organizational scheme, you can create labels for the cabling and equipment in your system that provide a great deal of information. For example, a horizontal (work area) cable with the label "1A-B02" tells you that the origin point is the first floor, TS A, patch panel B, position 02. A backbone cable with the label "1A/2A-1" tells you that this is cable 1, connecting TS A on the first floor to TS A on the second floor.

By uniquely identifying the cables and labeling them as they are installed, you provide the information required to manage the cabling system in your network. A technician working on the office end of a connection can easily locate the cabling closet that each cable comes from. The technician can then perform network tests or make changes to the system without having to spend any time tracing cables or disrupting other users while hunting for the right cable.

There are several companies that sell cable labels and printing software or standalone printers for the labels. These tools help to automate the labeling process in a large cable plant. The resources listed in Appendix A provide some pointers for locating suppliers of cable labels.

Documenting the Cabling System

An essential feature of a cabling system is documentation. For small networks that cover a single floor or a small set of floors, you can get by with an annotated copy of the building floor plan. While building your network, you should draw each cable installation on your copy of the floor plan and identify it. A separate notebook or a spreadsheet with an explanation of your cable identification system should be kept as well.

Cables should be identified when they are installed, and a label should be attached to each end of the cable, or onto the faceplate where the cable is terminated. After you've done this, an entry should be made in the cabling notebook or spreadsheet. It takes some discipline to ensure that this is done each time a network installation is made. However, over time you can create a document that will be quite valuable when it comes time to redesign the network, add new connections, or troubleshoot the system.

For larger systems covering entire buildings or sets of buildings, you may wish to consider a commercial software package designed to help manage cables. There are several packages designed for managing telecommunications cabling. Some of these packages are based on computer-aided design (CAD) software, and may include a database for handling the thousands of entries that a large cabling system can generate. Such packages are typically expensive and take a fair amount of time to set up and use. Nevertheless, if you are trying to manage large amounts of cable, a good cable management software package may be the only thing that can keep the system under some semblance of control.

While all of these systems may seem like extra work, they are really essential tools for your network. Unlike most tools, a structured cabling system and adequate documentation can easily be overlooked in the rush to design and install a network. However, by providing a cabling plan and documentation of your cabling system you will create a powerful tool for network management, and one that will pay real dividends when it comes to managing and troubleshooting the network.

Building the Cabling System

Once you've decided to install a structured cabling system, your next decision to make is: who will build the system? The logical choices are to build it yourself or hire a professional contractor. Which approach you choose will depend on the size and complexity of your cabling system, your budget, and the hardware skills of your staff.

At one extreme, a twisted-pair cabling system for a small workgroup is easy to set up and run. Twisted-pair Ethernet components can be very easy to work with. You can buy a small switch and some ready-made twisted-pair patch cables with eight-position plugs already connected. The patch cables are used to hook the stations to the switch ports, resulting in a complete network. If you need to install twisted-pair cabling for a whole floor or an entire building, however, things get more complex.

For large cable installations, there are several arguments in favor of hiring a professional cabling contractor. For one thing, a large cabling design can bring up a number of issues that you may never have heard of before. Cables that are installed in office buildings and other public spaces must meet a variety of stringent building safety and fire codes. To meet these requirements, many people prefer to hire a contractor who will see to it that things are done correctly. This may include installing new conduits, and making sure that any holes drilled through walls and floors are filled with fire-retardant material. A contractor who specializes in dealing with these issues can ensure that compliance with building standards and cabling standards is maintained when the cabling is installed.

Cabling System Challenges

There can be other, more critical issues as well when trying to install your network. For example, a network being installed in an older building may face the challenge of dealing with asbestos. In the United States and other countries, there are strict regulations in place that apply when disturbing these materials in any way. Regulations on these and other topics vary from state to state, and from country to country worldwide. Professional contractors can bring the expertise needed to deal with the special problems in your building, and with the regulations that may apply to your site.

Even if you decide to hire a cabling contractor to do the design and installation, you still need to make sure that the contractor knows exactly what your needs are. You also need to make sure that a careful cabling plan is followed, so that the cable installation is well documented and expandable in the future. The contractor should also test each installed cable and provide certification that the cabling meets the rated performance specifications. Some contractors will deal with these issues for you automatically, and some won't.

A cabling system for an entire building is a major project. When cabling a building, it is strongly recommended that professional cabling contractors be used. A cabling contractor knows how to design cable layouts for buildings and how to estimate cabling and installation costs, and should be able to help you with the planning for your system. A contractor can also evaluate your site for special problems, and develop estimates for asbestos abatement if required.

When evaluating a cabling contractor, you can ask the contractor for the names of previous customers. You can then ask the customers whether the contractor completed

the job on time and within budget, and whether they were happy with the resulting cabling system.[2]

For a smaller design involving a limited area, you may decide to forgo using outside contractors. This assumes that you have access to technicians with the appropriate training, skills, and tools, or that you're willing to install and test the cables yourself. Chapters 16 and 17 provide more details on the cables themselves. Note, however, that while these chapters provide guidelines and instructions on how to build horizontal cable segments, they do not describe how to install an entire building-wide cabling system.

2. The Building Industry Consultants Service International (BICSI) offers an "ITS Design Fundamentals Program" that includes a series of courses on cabling system fundamentals, cabling project management, etc.

CHAPTER 16

Twisted-Pair Cables and Connectors

The cables and components used to build a twisted-pair horizontal cable segment are based on the ANSI/TIA-568-C structured cabling specifications, which are designed to support all twisted-pair Ethernet media systems. These specifications are described in Chapter 15.

In this chapter, we'll show how a twisted-pair cable segment is wired and describe the components that are typically used. You may find this useful even if you don't build your own horizontal cable segment. Knowing the components and wiring standards used in cable segments can help you make sure that your cabling system is assembled properly. Being able to find your way around a cabling system is also a major benefit when it comes to troubleshooting network problems.

Following the sections on the various components of a twisted-pair cable segment, you'll learn how to install an RJ45 connector on a twisted-pair patch cable. The chapter concludes with special twisted-pair cabling considerations for the three twisted-pair Ethernet media systems, including the signal crossover wiring required by each system.

Horizontal Cable Segment Components

A horizontal cable segment is one that travels from a wiring closet to a work area, connecting an Ethernet switch with a station. This is the most widely used cable segment type in a structured cabling system. Building a horizontal twisted-pair segment involves the following set of components and specifications:

- Twisted-pair cable
- Eight-position connector
- Four-pair wiring schemes
- Modular patch panel used to hold eight-position jacks

- Work area wall outlet
- Twisted-pair patch cables and equipment cables with eight-position plugs

We will look at each of these items in turn and see how they can be used to build a twisted-pair cable segment

Telephone Industry Jargon for Wire Termination

The components and techniques used in structured cabling systems were first developed in the telephone industry. In the telephone industry, to *terminate* a wire means to attach the wire to a connector or wiring panel of some sort. A wire termination panel is a set of connectors to which the eight wires in a four-pair twisted-pair cable are attached.

Wire termination devices are widely used in the telephone industry and include patch panels, cross-connect blocks (also called punch-down blocks), and cable jacks and plugs.

Twisted-Pair Cables

Twisted-pair copper cable is quite different from the thick or thin coaxial cables used in the original Ethernet media systems, and the twinaxial cables used in 10, 40, and 100 Gb/s short reach Ethernet segments. The major difference is that the electrical characteristics of twisted-pair cable are not as tightly controlled as they are with coaxial cable. This makes transmitting high-frequency electrical signals over twisted-pair cabling a more difficult engineering task, because the signals have to deal with a harsher electrical environment. That, in turn, is why twisted-pair segment lengths are limited to a maximum of 100 m.

The twisted-pair cable specified for building a horizontal link consists of a set of solid wires with a thickness of between 26 and 22 AWG, surrounded by a thin layer of insulation. A 22 AWG wire has a diameter of 0.644 mm (0.0253 inches), and a 26 AWG wire diameter of 0.405 mm (0.0159 inches). The thin, solid wire is low-cost, and it's easy to install the individual wires in the punch-down connectors widely used for terminating wires in structured cabling systems. This type of connector is also called an *insulation displacement connector* (IDC).

An IDC allows a solid wire to be "punched down" into the connector without stripping off the insulation. Instead, the sharp edges of the connector components displace the insulation and grip the metal core of the wire as it is pushed into the connector with a punch-down tool. The punch-down tool cuts off any excess wire at the same time that it punches the wire down, making the task of attaching twisted-pair wires to connectors quick and easy.

Figure 16-1 shows an office outlet with two eight-position RJ45-style jacks using 110-type punch-down wire terminators. The 110-type wire terminators are widely used, and the ones shown here are low-density single connectors, which makes it is easier to see how insulation displacement works.

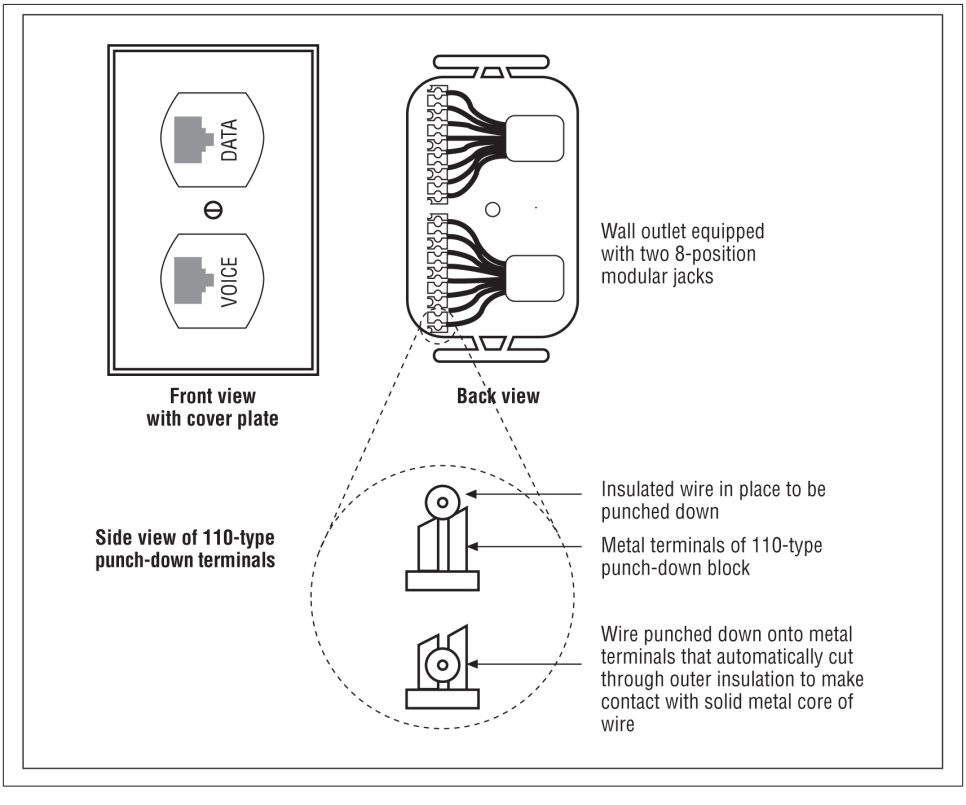


Figure 16-1. Punch-down connector

To illustrate how a wire is terminated using this type of punch-down block, a side view of a single 110-type wire terminator is shown in exaggerated scale below the office outlet. The twisted-pair wire is placed in the jaws of the metal terminals, and a punch-down tool is used to force the wire into the narrow space between the metal terminals of the wire terminator. The metal terminals automatically cut through the outer insulation of the wire and tightly grip the solid metal core, providing an electrical connection to the wire.

There are a variety of punch-down blocks and other kinds of wire terminators available in the cabling industry; they all use this same basic technique for making wire connections. Newer wire termination systems that are rated for Category 5e and 6A operation use a variety of IDC assemblies that are designed to hold the wire pairs in the correct

orientation to one another, and to terminate the wires without a significant amount of wire being untwisted. These newer designs for IDC connectors help to maintain the signal quality.

The correct installation of Category 5e and 6A cables requires the use of the IDCs provided by the vendor for its plugs and jacks, and also requires that you carefully follow the vendor's installation instructions.

Twisted-Pair Cable Signal Crosstalk

When it comes to transmitting signals over twisted-pair cable, one of the most important cable characteristics is signal crosstalk. Signal crosstalk occurs when the signals in one wire are electromagnetically coupled, or crossed over, into another wire. This happens because wires in close proximity to one another can pick up each other's signal. In a twisted-pair Ethernet segment, excessive crosstalk can result in the signals from the transmit wires being coupled into the receive wires. This increases electrical noise levels and signal error rates, and can also cause problems with collision detection on segments operating in half-duplex mode.

The way to avoid excessive crosstalk is to use the correct type of twisted-pair cable, and to ensure that each pair of wires in a twisted-pair segment is twisted together for the entire length of the segment. Twisting the two wires of a wire pair together minimizes the effect of electromagnetic signal coupling between pairs of wire in the cable. This helps make sure that any interference between the wire pairs is below the level of crosstalk that the twisted-pair transceivers are designed to ignore.

Twisted-Pair Cable Construction

One major difference in construction between the various categories of twisted-pair cable has to do with how many twists per foot the wire pairs have been given. The wire pairs in a voice-grade Category 3 cable typically have two twists per foot. This is lightly twisted wire, and you may have to strip back a good bit of outer insulation on a Category 3 cable to reveal the twists.

The wire pairs in higher-grade cables have progressively more twists per foot. Category 5e and 6A wire pairs are tightly twisted, which results in improved crosstalk performance at higher frequencies. For instance, one vendor specifies the wire pairs in a Category 5e cable as having from 19 to 25 twists per foot. Individual pairs within the cable have a different number of twists, which helps reduce the amount of signal that transfers between pairs.

Another characteristic of twisted-pair cables is the type of insulation used on the wires and the cable jacket. Plenum-rated insulation is more stable at high temperatures and provides superior electrical characteristics. Standard PVC insulation will perform as rated for normal room temperatures, but at temperatures above 40°C (104°F) the signal

attenuation of PVC-insulated cable increases markedly. Therefore, plenum-rated cables provide better temperature stability and help ensure that the signal quality of your cabling system will remain high. A form of Teflon® called *fluorinated ethylene propylene* (FEP) is often used for the outer jacket of plenum cables. FEP, the most common form of Teflon, is also used as insulation on the individual wires inside cables to improve signal quality and stability.

Plenum-rated cables are typically required for installation in building air-handling spaces (also called *plenums*) to meet fire regulations. The reason for this is that different kinds of plastic cable insulation behave differently in a fire. PVC insulation is "fire retardant" in comparison to plain polyethylene plastic, but PVC will still burn and produce smoke and heat. Teflon FEP insulation produces much less smoke and heat when burning, and does not support the spread of flames.

Plenum cable identifiers

The *National Electric Code* (NEC) provides identifiers for communications wires and cables:

CMP
: Cables with a CMP identifier are plenum-rated and are suitable for installation in ducts and plenums without the use of conduit. These cables are designed for fire resistance and low smoke-producing characteristics.

CMR
: Cables with a CMR identifier are not plenum-rated. However, they are engineered to prevent the spread of fire from floor to floor and are suitable for riser use and vertical shaft applications.

CM
: Cables with a CM identifier are specified for general-purpose building wiring use, in areas other than plenums and risers.

By looking on the cable for these cable marks, you can tell if a particular cable is suited for a given installation. There is no major difference between CM and CMR cables, because they are both based on PVC insulation; CMR cables simply have more fire-retardant material in them, to help slow down the spread of flames.

Shielded and unshielded twisted-pair cable

The majority of twisted-pair cable installed in the United States is unshielded. The cabling industry has developed several versions of shielded twisted-pair cable, which has improved signal-carrying characteristics. However, shielded twisted-pair cables are also more expensive, and require correct installation in properly grounded equipment so as to maintain proper grounding of the shield to avoid signal impairment due to electrical issues caused by incorrect grounding. Cabling standards in countries other than the

United States have specified shielded cable to meet various regulations, with the result that shielded cable is more widely used in Europe.

The descriptions used for shielded twisted-pair cable have evolved over time, leading to some confusion. The terms include:

Screened twisted pair (ScTP or F/TP)
: ScTP cabling is provided with a *single* foil or braided screen (shield) across all four pairs within the twisted-pair cable, which minimizes electromagnetic (EMI) radiation and susceptibility to electrical noise from outside the cable. The designation F/TP means that the cable uses foil shielding instead of a braided screen.

Screened shielded twisted pair (S/STP or S/FTP)
: S/STP (screened shielded twisted pair) or S/FTP (screened/foiled twisted pair) cabling provides shielding between the pairs *and* an overall shield around all twisted pairs within the cable. This type of shielding minimizes the level of EMI that can enter or exit the cabling, and also minimizes crosstalk between neighboring pairs within the cable.

Shielded foiled twisted pair (SFTP)
: SFTP has both foil and braided wire shield together, located around all four wire pairs.

Shielded twisted-pair naming conventions

The naming conventions in the ISO/IEC 11801 standard describe two types of shield:

Overall shield
: This shield is located around all of the twisted pairs together, and can be one of three types:

  - F = foil shielded
  - S = braided-wire shielded
  - SF = braid and foil together

Element shield
: This shield is on each twisted-wire pair inside the cable, and can be one of two types:

  - U = unshielded
  - F = foil shielded

In the ISO standard, the overall shield is used in the first part of the cable identifier and the element shield in the second, with the parts separated by a slash. As an example, S/FTP is braid shielded on the outside of all four wire pairs, with each wire pair foil shielded. The primary cable used in many European countries is S/FTP, rated by the ISO/IEC standard as Category 7 cabling.

It is expected that the new TIA Category 8 cable, whose specifications are under development by for use on 40GBASE-T links, will be an F/UTP cable, with a foil overall shield and four unshielded twisted pairs.

Twisted-Pair Installation Practices

Most structured cabling systems are installed by professional cabling contractors. These contractors have the expertise and equipment required to correctly and safely install the hundreds of twisted-pair cables that a typical office building can require. Cable contractors are also familiar with the structured cable standards and will ensure that the cabling systems that they install meet the specifications.

The currents and voltages used to carry Ethernet signals over twisted-pair wires are small and pose no threat to the users of Ethernet equipment. However, the twisted-pair wires used for telephone services or to power circuit repeaters used in high-speed data lines may carry large currents and voltages. Always observe standard safety practices when working on any type of wire, and take all necessary precautions to avoid electrical shock.

Should you decide to install a small twisted-pair cable system or a few horizontal segments, the ANSI/TIA/EIA standards provide cable installation guidelines. These guidelines are intended to minimize any effect on the wire twists inside the cable. To support high-speed signals, the wire twists in the cable must remain tightly twisted and not be disturbed anywhere along the length of the cable. Cable ties and fasteners that are too tight, or outer cable jackets that are excessively twisted, can affect the wire twists inside the cable. The installation guidelines include the following:

Maintain the minimum bending radius
: The minimum bending radius for a four-pair cable should be eight times the outside cable diameter. If the cable diameter is 0.5 cm (0.20 inches), the minimum bend radius will be about 4.0 cm (1.57 inches).

Minimize jacket twisting and compression
: Install cable ties loosely and use Velcro® fasteners that allow the cable bundle to move around a bit. Take precautions to avoid tightly compressing the jacket of the cable. Do not use staple guns to fasten the cable to backboards.

Avoid stretching the cable
: Do not exceed 110 newtons (25 pounds-force) of pulling tension when installing the cable.

Keep wire twists intact to within 13 mm (0.5 inches)
: This applies to wire termination for Category 5, 5e, and 6A systems. For example, when making a wire termination in an eight-position jack, do not untwist any further back than 0.5 inches from the end of the wire pairs in the cable.

Avoid close proximity to power cables or other electrical equipment
: A distance of 30.5 cm (12 inches) is recommended between horizontal cables and fluorescent lighting fixtures. A distance of 1.02 m (40 inches) is recommended for transformers and electrical motors.

  If the horizontal cable is in a metal conduit, then a distance of 6.4 cm (2.5 inches) is recommended for unshielded power lines carrying less than 2,000 volts. If the horizontal cable is in an open or nonmetal pathway, then a distance of 12.7 cm (5 inches) is recommended for unshielded power lines carrying less than 2,000 volts.

Eight-Position (RJ45-Style) Jack Connectors

The eight-position connector used in the ANSI/TIA-568 standards is formally described as one that meets the requirements specified in the IEC 603-7 standard for eight-way connectors. You will often hear the eight-position connector referred to as an *RJ45-style connector*, which is the name originally used by the telephone industry. The RJ45 name comes from *Registered Jack*, which was an official U.S. telephone industry designation for an eight-position connector.

At one time, the U.S. telephone service was a monopoly, which organized its operation in terms of services that were registered with various public utilities commissions. The specifications for these services include such things as the jack connectors used to provide wire termination for the services, hence the name *Registered Jack*.

To make sure that the entire segment can carry high-frequency signals without excessive signal distortion, crosstalk, or signal loss, *all* of the connection components in the horizontal channel must be correctly installed and rated to meet the category specifications for the cabling involved.

For a Category 5e cabling system, simply installing Category 5e cable is not enough; all of the other components used in the segment must also meet the Category 5e specifications. Standard telephone-type voice-grade RJ45 connectors are widely available, but they do not meet the Category 5e specifications. Instead, to provide a segment that meets the Category 5e specifications, you must be sure to use eight-position connectors and other components that are specifically designed for use in Category 5e cable systems.

Four-Pair Wiring Schemes

For a horizontal cable segment, the ANSI/TIA-568 standards recommend the use of four-pair cables with all eight wires terminated in eight-position jack connectors at each end of the link. The entire twisted-pair cabling system should be wired "straight through." This means that pin 1 of the connector at one end of a horizontal cable is wired to pin 1 of the connector at the other end, and so on for all eight connections. This keeps the structured cabling system very simple and straightforward.

Tip and Ring

The words *tip* and *ring* are used to identify wires in a wire pair. Most single analog telephone circuits require just two wires to deliver what is known in the telephone industry as *plain old telephone service* (POTS). These two wires are identified as "tip" and "ring" by the industry. These names date from the earliest days of manual telephone switchboards, when operators made connections between telephone lines using patch cables with plugs on the end.

The plugs had a tip and a ring conductor on them; hence the names for the two wires still used to make a basic analog telephone connection. Each pair of wires in a modern communications cable is still considered to have a designated tip conductor and ring conductor, labeled T1 and R1 for the first pair, T2 and R2 for the second pair, and so on.

Color Codes

To help identify all the wires found in a multipair communications cable, the telephone industry developed a widely used system of color coding. This system uses a pair of colors to identify the individual wires in each wire pair. The primary color group consists of white, red, black, yellow, and violet. The secondary group uses the colors blue, orange, green, brown, and slate. These colors are used to identify the wires in the majority of twisted-pair communications cables, from two-pair cables on up to larger cables.

A primary color is paired with one of the secondary colors for each wire in the cable. For large cables, the primary color is used until it has been combined with each of the five secondary colors. Then the next primary color is paired with each of the five secondary colors, and so on. In a typical four-pair cable, the primary color is white, and no other primary color is needed because there are only four pairs.

Starting with the first wire in the first wire pair of a cable (T1), the insulation is given a base coat of the first primary color, white, with a stripe or dash of the secondary color blue. This is written as "white/blue" and is abbreviated as W-BL. The second wire in the first wire pair (R1) is given a base coat of the secondary color, blue, with a stripe or dash of the primary color white, written as "blue/white" and abbreviated as BL-W (or sometimes just BL). In the first wire pair, then, the T1 wire is white with a blue stripe, and

the R1 wire is blue with a white stripe. In the second wire pair, wire T2 is white with an orange stripe, and R2 is orange with a white stripe, and so on.

Wiring Sequence

The term *wiring sequence* refers to the order in which the wires are terminated on a connector. There are two wiring sequence options provided in the ANSI/TIA-568 standards. The *preferred* wiring sequence according to the standards is called T568A, and the *optional* wiring sequence is called T568B.

Figure 16-2 shows the preferred and optional wiring sequences for an eight-position jack connector. Which sequence you use is a local decision. Note that the words "preferred" and "optional" may not reflect reality at your site.

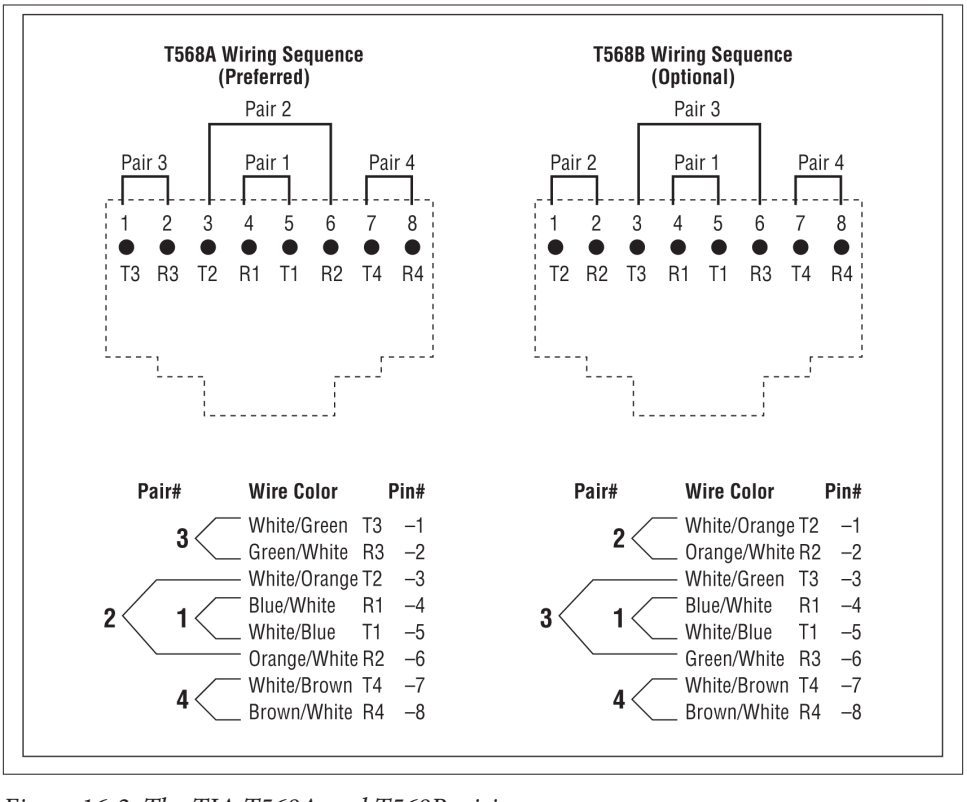


Figure 16-2. The TIA T568A and T568B wiring sequences

The optional wiring sequence is widely used, and many cable installers use it as their default cabling standard. That's because the optional T568B sequence is also known as the AT&T 258A wiring sequence, and has been widely used for years in AT&T cabling

systems. It's up to you to find out what wiring sequence is widely used at your site, and to make sure that your cabling system adheres to the local standard to avoid confusion.

The center two positions in both wiring sequences, pins 4 and 5, are always used for pair 1—this is where telephone voice circuits are wired if the link is used for analog voice service. That's why the 10BASE-T standard originally specified the use of positions 1, 2, 3, and 6, avoiding the use of pins 4 and 5: that way you could run a 10BASE-T service and analog voice service over the same four-pair cable if you wished. Although most installations preferred to keep the analog voice and data services on separate cables to avoid the problem of noise from telephone ringing circuits affecting the data service, subsequent Ethernet twisted-pair media standards based on two pairs followed the 10BASE-T wiring scheme for cabling compatibility.

Keeping the wires correctly paired together for the entire length of the horizontal channel is critically important to maintaining signal quality for Ethernet signals. As it happens, there is an older wiring sequence that you may encounter in existing cabling systems that does not provide the correct wire pairing and that can lead to problems for Ethernet signals. The older wiring sequence that results in incorrect wire pairing is called the Universal Service Order Code system (USOC). Despite the name, this is not a universally adopted system, but it was used in older telephone systems. The USOC system deals with the pairs differently, and the wire identification used in the old USOC system is often based on an older color scheme as well.

Because of the way the pairs are wired in the USOC scheme, you will end up with a *split pair* if you try to install a twisted-pair Ethernet segment on a cable using the USOC wiring sequence.

Figure 16-3 shows an eight-position connector with USOC wiring based on the older color-coding scheme that is frequently used with USOC systems. For comparison, the T568A wiring sequences is also shown. Notice that the wires connected to pins 1 and 2, which are paired together in both the T568A and T568B wiring sequence, are not paired together in the USOC scheme. If you plug a twisted-pair Ethernet station into a cabling system that is wired using the USOC scheme, the Ethernet segment can end up with excessive signal noise and crosstalk because of the split-pair wiring.

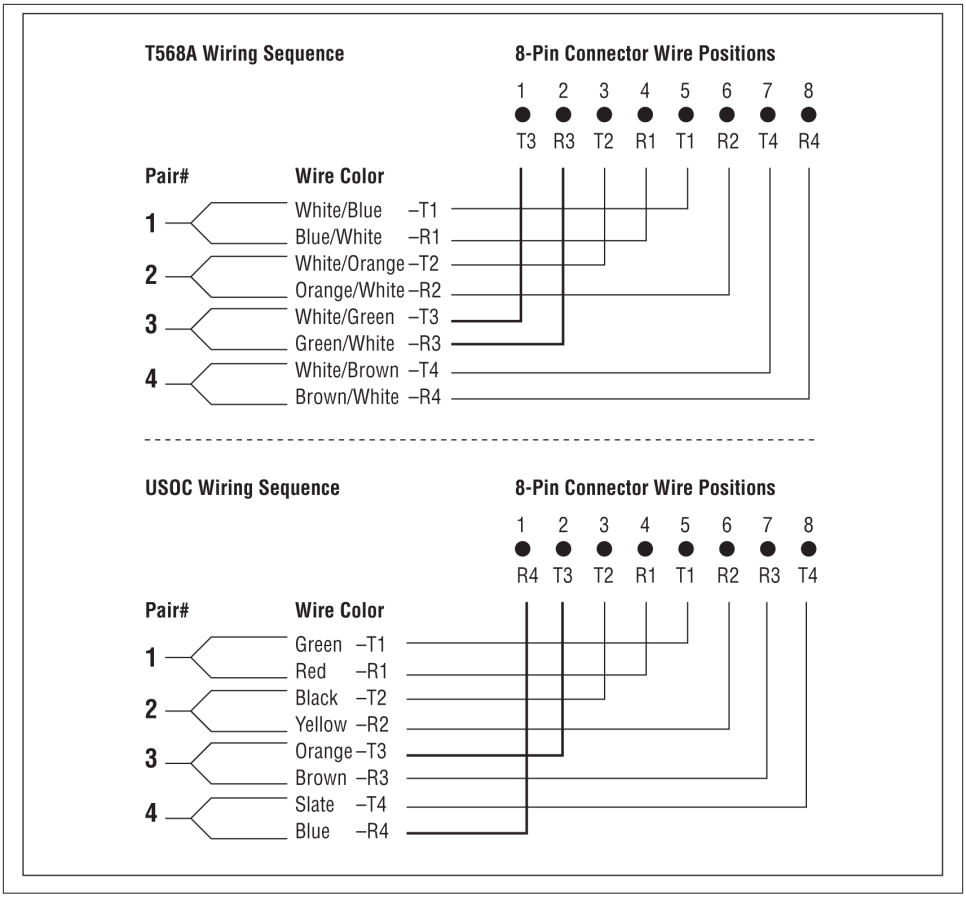


Figure 16-3. Split pairs in USOC wiring

It might not be immediately obvious that there's a problem with the segment, because a simple wiring test of the connection would show that basic wire connectivity from end to end is OK. In other words, the USOC wiring sequence provides wires between every pin of the eight-pin connectors located at each end of the link, so that checking for connectivity between the pins at each end of the link will not detect the problem. What USOC doesn't provide is the *correct pairing* of the wires on the wire pairs used to carry Ethernet signals.

It may seem odd that just twisting the wires together in a pair would make this much difference, but it does. Ethernet signals operate at high frequencies, where the lack of twists on a pair of wires makes a big difference in the electrical characteristics of those wires. If the correct wires are not twisted together for the full length of the segment, the segment will experience excessive signal noise and crosstalk, and may fail to operate properly.

Modular Patch Panels

Modular patch panels are panels designed to hold a number of RJ45-style jack connectors. The eight wires of the horizontal link cable are terminated in the jack connector, and the connector is installed in the patch panel, which is located in the telecommunications closet. You then use a patch cable to connect the jack in the patch panel to another patch panel or to hub equipment located in the closet, depending on how your cabling system is organized. You can buy patch panels that come fully populated with connectors, or you can get blank panels and simply add the number of connectors you need.

Figure 16-4 shows a modular patch panel of the sort used in telecommunications closets. From the patch panel, a horizontal link cable travels to the work area wall outlet, where the link cable is terminated in an eight-position modular jack. A patch cable is shown connected to the work area outlet. The other end of the patch cable could be connected to a computer in the office.

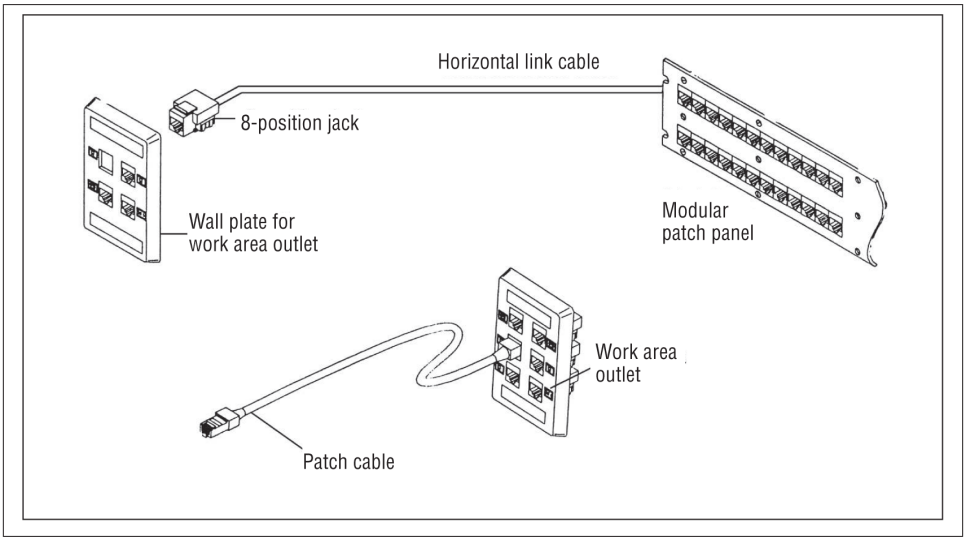


Figure 16-4. Modular patch panel, work area outlet, and patch cable

Modular patch panels provide a great deal of flexibility. You can also use several different patch panels in a given wiring closet, dedicated to different services. When you add new network equipment, you can provide separate patch panels for that equipment and easily connect different offices to different network equipment in the wiring closet, depending on what the user requires.

Work Area Outlets

The eight wires of each horizontal link cable are terminated in a modular eight-position jack connector mounted in a wall plate in the office space or work area. The telephone industry has had years of experience in wiring offices, and consequently there are a wide variety of wall plates available for terminating twisted-pair cables.

You can buy wall plates that range from a fixed pair of simple eight-position jacks to a more complex modular system that allows a wide variety of connectors to be installed in the same wall plate. A modular wall outlet makes it an easy task to provide a neat, low-cost, and reliable office connection to the horizontal cabling system.

Twisted-Pair Patch Cables

Each end of the horizontal link is connected to equipment such as Ethernet switches or computers using patch and equipment cables. At the wiring closet end of the link, patch and equipment cables are used, for example, to connect the link to an Ethernet switch, or into the backbone cabling system. At the work area end of the link, a patch cable is used to make the connection between the computer in the office and the jack in the wall outlet.

Patch cables must be very flexible to allow lots of movement, and for that reason they must use stranded twisted-pair wire instead of the solid kind. If you repeatedly bend solid conductor cable, the solid conductor inside the insulation of the cable will eventually crack and break, which will result in an intermittent failure that can be very hard to track down. Stranded cable, on the other hand, can withstand large amounts of bending and twisting without problems.

Twisted-Pair Patch Cable Quality

You can buy patch cables from cable suppliers at reasonable cost. Because it's easy to buy ready-made patch cables, many sites choose to avoid the problems with building their own patch cables by purchasing them from suppliers. This also takes advantage of the fact that a good-quality manufactured patch cable will be built using the correct connectors and stranded cable and according to standardized manufacturing and test procedures.

While it's not impossible to build a good homemade Category 5e cable, it takes a lot of attention to detail, and it may cost more than you might expect to do the job correctly. For example, while all RJ45 connectors may look alike at first glance, there are small differences in the way they are built and the way they fit into a crimping tool. Finding the exact match between the crimping tool and the connector you use can be difficult.

There are a lot of RJ45 crimping tools on the market, too, and a number of them are of low quality, with flimsy plastic or lightweight metal frames that may not provide enough

force to produce a really solid crimp. High-quality crimping tools are expensive, and often require special crimping dies designed for vendor-specific versions of RJ45 connectors.

Therefore, it's a good idea to buy high-quality patch cables ready-made from a reputable manufacturer. This is especially true for any system that will be supporting higher-speed Ethernets, such as Gigabit and 10 Gigabit Ethernet, which send signals over all four wire pairs simultaneously. Also, 10 Gigabit Ethernet requires Category 6A cables, which are more difficult to build correctly. Maintaining the best possible signal quality for 10 Gigabit Ethernet is critically important to the operation of the link, which argues against using anything but the best available manufactured cables.

There are a lot of patch cables on the market, so you need to make sure that the patch cables you buy are rated to meet Category 5e or 6A specifications, to match your cabling system. Very low-cost or generic patch cables may not be carefully built with quality components and may not meet the specifications or maintain their rating over time.

Telephone-Grade Patch Cables

Beware of using standard telephone-grade patch cables for twisted-pair segments. One common patch cable used in the telephone industry goes by the name of *silver satin*, which describes the outside color of the cable. This is the flat patch cable that is often used to connect an analog telephone to a wall jack, and this cable is widely stocked in ordinary hardware or office supply stores.

The biggest problem with this type of patch cable is that the conductors in silver satin cords are not twisted together, leading to excessive levels of crosstalk on the wires in this cable. This can potentially cause spurious frame errors on your segment. Another problem with silver satin is that the conductors are quite small, which causes higher signal attenuation. Therefore, using silver satin significantly reduces the distance that a signal may travel.

One of the worst problems with silver satin cable is that despite all the signal errors, it may work OK at the lowest supported speed (10BASE-T) when it is used in an Ethernet segment. However, the silver satin patch cable may still be causing data errors and lost frames. These problems can be masked because the Ethernet system will keep trying to function despite the errors, and the problems on a single segment may not cause the rest of the network to fail.

That, coupled with the fact that each station's high-level protocol software will keep retransmitting frames until something gets through, tends to hide the effects of a poorly functioning media system. However, the higher the traffic rate gets, the more these errors will occur, often leading to complaints of a slow network.

As things progressively get worse, you will be forced to find all of the silver satin patch cables and replace them with the right kind of twisted-pair patch cable. A better ap-

proach is to simply forbid the use of any wire or other component in a horizontal cabling system that does not meet the category specifications for your cabling system (e.g., Category 5e or 6A). Also, make sure everyone understands that silver satin patch cables are something that must be avoided in any structured cabling system designed to carry data signals.

Twisted-Pair Ethernet and Telephone Signals

A twisted-pair Ethernet transceiver is often attached to a twisted-pair segment with a patch cord connected to an RJ45-style modular jack in a wall outlet. One RJ45 modular jack looks a lot like another, and you can mistakenly connect a transceiver to a telephone outlet instead of the correct data outlet.

The center two pins of the RJ45 jack (pins 4 and 5) may be used by analog telephone services. Therefore, to avoid a conflict with telephone services, the 10BASE-T and 100BASE-T systems do not use pins 4 and 5. These days, however, all Ethernet segments are wired with all four pairs, as required to support 1000BASE-T and faster Ethernet. This makes it much likelier that these Ethernet cables, when mistakenly installed in a telephone jack, could receive analog telephone signals.

The telephone battery voltage is generally 56 VDC, and telephone ringing voltages include an AC signal of up to 175 V peak with large transient voltages at the start and end of each ring interval. Thus, there is a possibility that an Ethernet transceiver in a device could be damaged by these voltages. The standard notes that while Ethernet equipment is not required to survive such wiring hazards without damage, the equipment manufacturers must ensure that there will be no safety hazard to the user from the telephone voltages.

According to the standard, an Ethernet transceiver typically appears as an *off-hook* telephone to the analog telephone system, meaning that the telephone is in use. Because the telephone system will not send ringing voltages to an off-hook telephone, this should help prevent any damage to an incorrectly connected Ethernet device.

Equipment Cables

In the telecommunications closet, the *equipment cable* is the cable that connects the active equipment, such as an Ethernet switch, to the patch panel. The equipment cable might be as simple as a patch cable, or may include cables that are more complex. For Ethernet switches with RJ45-style jacks on the front, you simply connect patch cables from each jack on the hub to the appropriate jack on the patch panel in the wiring closet, and you're done.

50-Pin Connectors and 25-Pair Cables

You may encounter older 10BASE-T Ethernet switches that are equipped with 50-pin connectors instead of RJ45-style jacks. This approach was used in older Ethernet switches when a manufacturer wanted to accommodate a large number of connections on a switch panel or modular card for a chassis switch.

In that case, a single 50-pin connector was used to provide 12 four-wire connections, allowing a vendor to support 24 connections on a single interface board with just two 50-pin connectors. The 50-pin connectors, and the 25-pair cables they connect to, were traditionally used in voice-grade cabling systems and were typically rated for Category 3 performance. Therefore, this approach was more popular back when 10BASE-T Ethernet was new. Note, however, that newer versions of these cables and connectors were developed that were rated for Category 5 use.

While using prewired 25-pair cables for connections to patch panels and switches can minimize the amount of wiring you have to do in a wiring closet, there are serious drawbacks. For one thing, they are limited in signal quality and cannot support higher-speed versions of Ethernet. Also, it can be more difficult to troubleshoot a network problem in this kind of installation, because there is no easy way to move a connection from port to port of the Ethernet switch. Because all connections are wired simultaneously with the 25-pair cable, you can't pull one connection out and try it on another hub port as a test, making it much more difficult to isolate a problem to a particular horizontal cable.

25-Pair Cable Harmonica Connectors

A cable *harmonica* is a small plastic housing equipped with a strip of RJ45-style jacks in a row, so named because the row of RJ45 holes on the housing makes it look somewhat like the musical instrument. The harmonica terminates one end of a 25-pair cable whose other end is equipped with a 50-pin connector for connection to an Ethernet switch. This system typically supports up to 12 RJ45-style jacks per harmonica.

Building a Twisted-Pair Patch Cable

The following is a quick reference guide to the installation of an RJ45 plug onto a patch cable. Twisted-pair patch cables should only be made using stranded wire cable. Solid wire cable is unacceptable for patch cables, because it will break when flexed, causing intermittent connections. If you choose to build your own patch cables, you need to buy stranded twisted-pair cable and the correct RJ45-style plugs for terminating stranded wire.

Because solid conductor cable is specified for use in the horizontal cable segment, many RJ45 connectors are designed for use on solid conductor cable and can cause problems when crimped onto a stranded patch cable. Using the wrong connector on a stranded

twisted-pair cable could cut too deeply into the conductors of the wire and weaken them so that they may break easily, which can result in an intermittent connection. To avoid this, you need to make sure that you are using RJ45 plugs that have been specifically designed for stranded wire.

Duplicating the carefully controlled manufacturing process yourself can be quite difficult. Without careful attention to a number of important issues, the result may be a cable with a connector that doesn't really fit the cable correctly, and that may have been crimped onto the cable with the incorrect tool. Although such a cable may initially pass when tested with a cable tester, these problems can eventually lead to intermittent connections and network outages.

Reputable cable and connector manufacturers employ engineers who ensure that all components and tooling used in the manufacturing process are correct, and that every connector is installed in a consistent manner. These engineers put samples of the manufactured cable assemblies through tests to ensure that critically important characteristics, such as pull strength and electrical resistance, are being correctly maintained.

The result of the manufacturing process is a cable that is correctly mated to the connector, and a connector that is correctly installed using the right tool with the correct amount of pressure.

Installing an RJ45 Plug

Building a patch cable involves installing RJ45 plug connectors on each end of a stranded cable. Here we describe the process of installing the RJ45 connectors.

Attaching cable connectors involves the use of very sharp knives for stripping cable insulation as well as crimping tools that can be dangerous to operate. Many crimping tools incorporate a ratchet mechanism that, once engaged, prevents the tool from being opened until it has first closed completely. Anything caught in the crimping tool, including your fingers, will be crushed.

Here are the steps to follow:

1. Carefully strip away a few inches of the outer insulation from the twisted-pair cable, revealing the individually insulated twisted-pair conductors inside. Each twisted-pair conductor consists of a set of thin stranded wires surrounded by insulation. *Do not cut into the insulation of the twisted-pair conductors.*
2. Orient the conductors according to the colors of the insulation.

3. Straighten out the twisted-pair conductors, arrange them as shown in Figure 16-5, and cut the conductors to a length of about 12 mm (0.5 inches). Leave the insulation in place on the individual twisted-pair conductors. Make sure that the conductors are all cut to the same length, providing a square end to the cut.

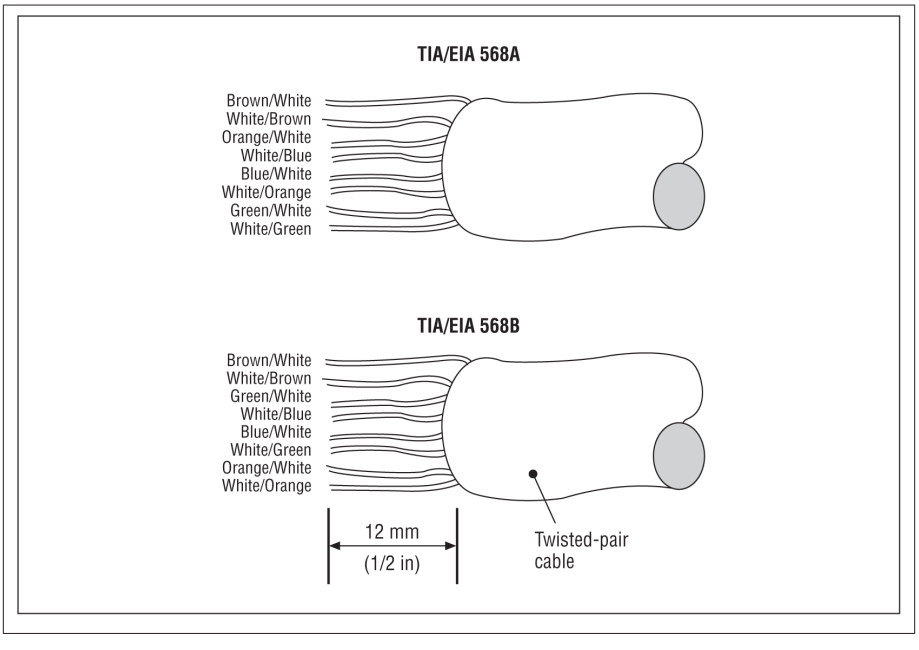


Figure 16-5. Arrange the twisted-pair wires

4. If you wish to use the TIA T568A preferred wiring sequence, arrange the conductors in the following order from top to bottom. The wire colors, and the pin numbers to which they connect, are as follows:

   - Pin 8: Brown/White
   - Pin 7: White/Brown
   - Pin 6: Orange/White
   - Pin 5: White/Blue
   - Pin 4: Blue/White
   - Pin 3: White/Orange
   - Pin 2: Green/White
   - Pin 1: White/Green

5. If you instead wish to use the TIA T568B optional wiring sequence (also known as the AT&T 258A wiring sequence), arrange the conductors in the following order from top to bottom:

   - Pin 8: Brown/White
   - Pin 7: White/Brown
   - Pin 6: Green/White
   - Pin 5: White/Blue
   - Pin 4: Blue/White
   - Pin 3: White/Green
   - Pin 2: Orange/White
   - Pin 1: White/Orange

6. Hold the RJ45 connector with the bottom (contact side) facing you. The blunt end of the connector (which gets inserted into an RJ45 jack) should be pointing to the left and the open end of the connector should point to the right.

   While holding the connector in this orientation, the pin 8 position is on the top edge, and the pin 1 position is on the bottom edge. Hold the twisted-pair cable firmly in your other hand. Insert the insulated twisted-pair conductors into the connector as shown in Figure 16-6. Make sure to keep the conductors in the correct sequence.

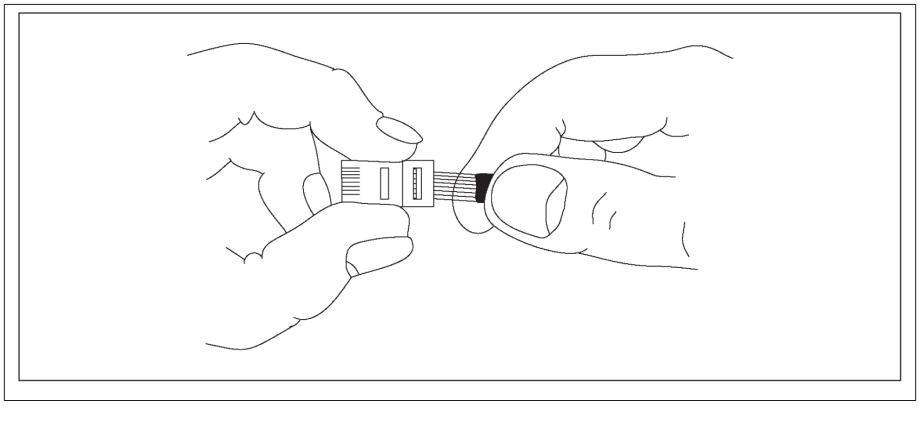


Figure 16-6. Insert the conductors into the connector

7. Slide the conductors all the way into the connector, so that they are firmly seated against the inside front of the connector shell. When the conductors are all the way into the connector, you should be able to see the ends of the conductors through

the front of the connector (see Figure 16-7). The outer insulation of the cable should be under the strain relief clamp.

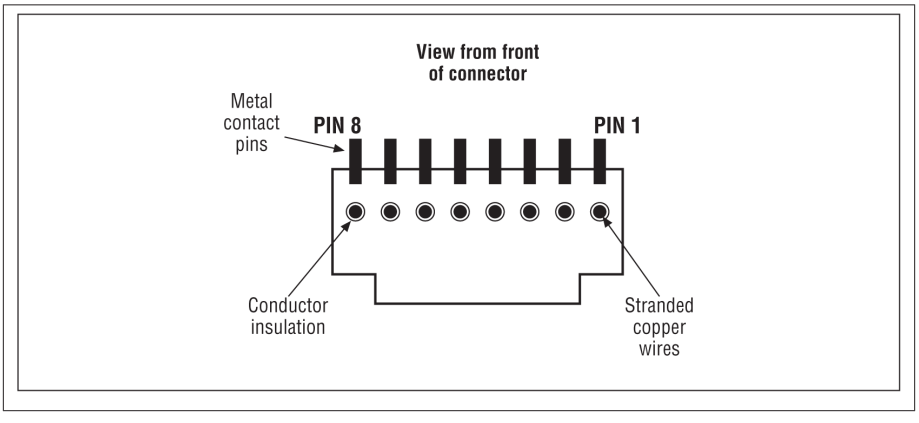


Figure 16-7. Conductors properly inserted inside the connector

8. While holding the cable and connector firmly together, insert them all the way into the crimping tool (see Figure 16-8). The connector will go all the way into the crimping tool only if it is inserted from the correct side. Before crimping, verify that the conductors are still properly seated inside the connector.

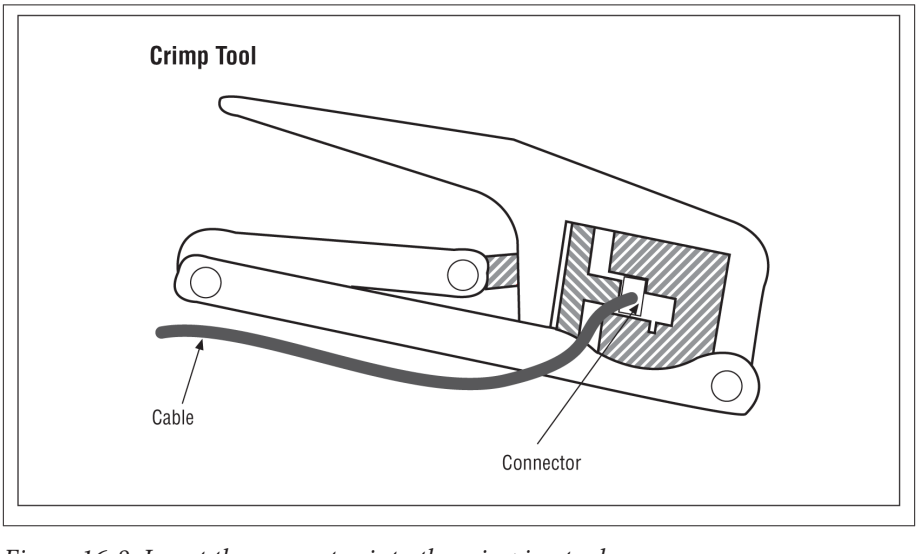


Figure 16-8. Insert the connector into the crimping tool

9. Place the flat base of the crimping tool on a solid surface, such as a table or floor. Press down the handle until it comes into contact with the stop. This forces the contacts inside the connector to bite through the insulation on the conductors. This also forces the cable strain relief assembly into place. The strain relief block is important, as it clamps the cable into place in the connector. This prevents stresses on the cable from pulling the conductors out of the connector.

   Figure 16-9 shows a connector before and after crimping. After crimping, the plug contacts bite through the insulation and into the copper wire portion of the twisted-pair conductors. The strain relief block is forced into place to hold the cable into the connector.

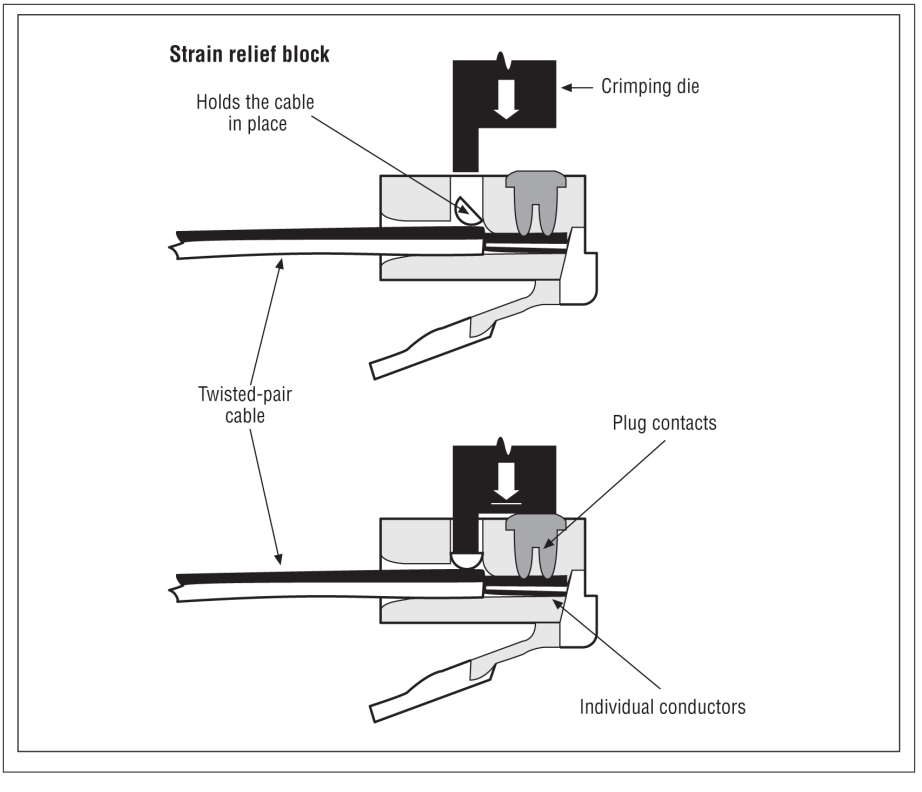


Figure 16-9. Connector before and after crimping

Ethernet Signal Crossover

To make the data flow when connecting two twisted-pair Ethernet transceivers together over a twisted-pair link segment, the transmit data signals of one transceiver must end

up on the receive data pins of the other transceiver, and vice versa. When the 10BASE-T and 100BASE-T standards were developed, the crossover wiring was accomplished in one of two ways: with a crossover cable, or by crossing over the signals inside the switch port, as shown in Figure 16-10.

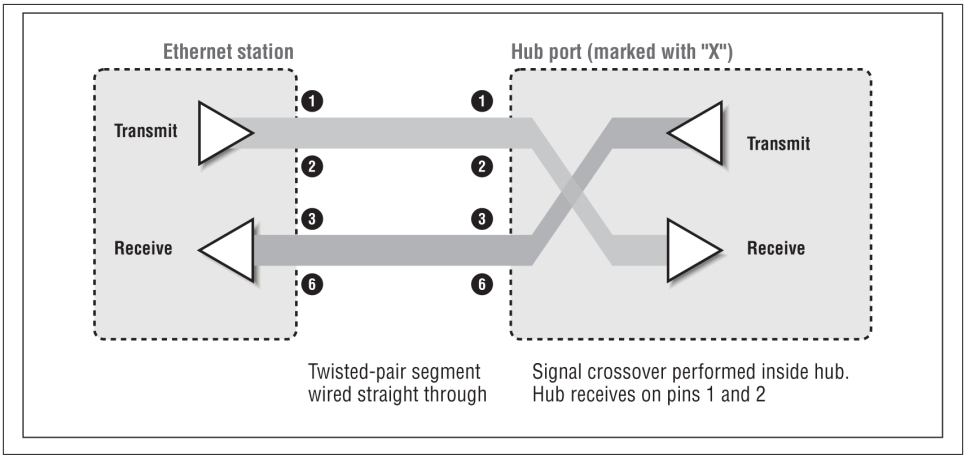


Figure 16-10. Signal crossover inside a switch port

The 1000BASE-T standard was accompanied by the development of the MDI/MDI-X auto-crossover specification, defined in Clause 40 of the standard. MDI-X, also known as MDIX and automatic MDI/MDI-X, was first developed in 1998; it is specified as an optional capability for the medium attachment unit (MAU) that "is intended to eliminate the need for crossover cables between similar devices."[1]

Today, most Ethernet interfaces in devices and switch ports provide automatic signal crossover, and if they don't they will usually provide a "hardwired" internal crossover. This relieves you of the task of supplying a crossover cable in the cabling system. Instead, each twisted-pair segment can be wired straight through, as recommended in the structured cabling standards.

10BASE-T and 100BASE-T Crossover Cables

In the fairly unusual case of networking only two devices, the two Ethernet stations can be linked together with a single cable. This eliminates the need for an Ethernet switch, but also eliminates the signal crossover that is done inside the switch ports. However, if the Ethernet interface in either or both of the devices implements Auto-MDIX, then they will automatically create a signal crossover when connected together. If the two

1. IEEE Std 802.3-2012, paragraph 40.4.4, p.227.

devices do not have the Auto-MDIX option, then you will need to build a crossover cable to make the signals work properly.

Another use for a crossover cable arises when you need to link switch ports together between two older switches that have hardwired signal crossover inside their ports, and that do not support Auto-MDIX. In this case, there are one too many signal crossovers being done, and this connection will not work with a straight-through cable. Therefore, you need to use a crossover cable to link the two ports.

Figure 16-11 shows the crossover wiring required for the original 10BASE-T and 100BASE-T systems, prior to the widespread adoption of Auto-MDIX. Because both of these media systems use the same four wires, a crossover patch cable or a switch port with internal crossover wired in this fashion works on both systems. Modern Ethernet interfaces almost all support Auto-MDIX, and there is generally no need to build special crossover cables anymore.

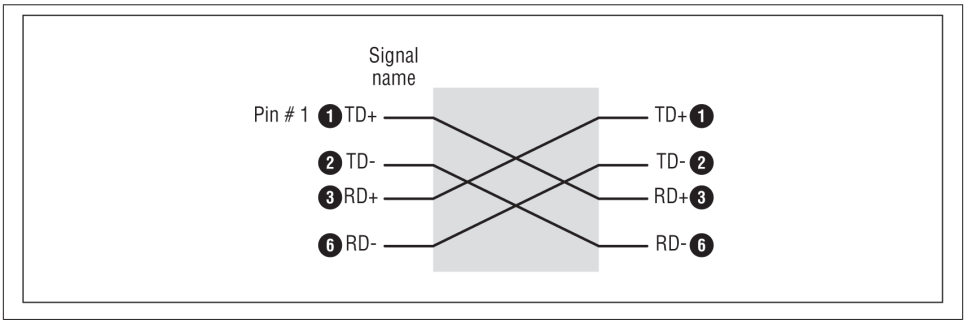


Figure 16-11. 10BASE-T and 100BASE-T crossover cable wiring

Four-Pair Crossover Cables

The Gigabit and 10 Gigabit Ethernet systems use all four pairs of wires, and require that all four wires be crossed over correctly to operate. To make this easier, the Gigabit Ethernet standard uses the MDI-X automatic crossover function, which is supported in most modern Ethernet transceivers.

In the automatic crossover system, the transceiver automatically moves the link signals to the correct logic gates inside the transceiver chip. Once a transceiver has moved the signals to different gates, it waits for approximately 60 milliseconds while checking the link for link pulses or data. This provides a mechanism for each end of the link to automatically configure the crossover function as needed. A random startup time is used to ensure that the ends of the link will not start moving the signals in synchronization, and thereby never achieve a correct crossover.

If neither or both of the ports you are connecting implement an internal crossover, then you can provide an external crossover to make the link work. You can provide the signal crossover for 1000BASE-T or 10GBASE-T links by building a crossover patch cable as shown in Figure 16-12. This crossover cable is universal, and will work for all other Ethernet twisted-pair media systems as well. Given that modern cabling systems provide all four pairs on each horizontal link, a four-pair crossover cable is the only kind of crossover cable you need to keep on hand for those increasingly rare instances when they are needed.

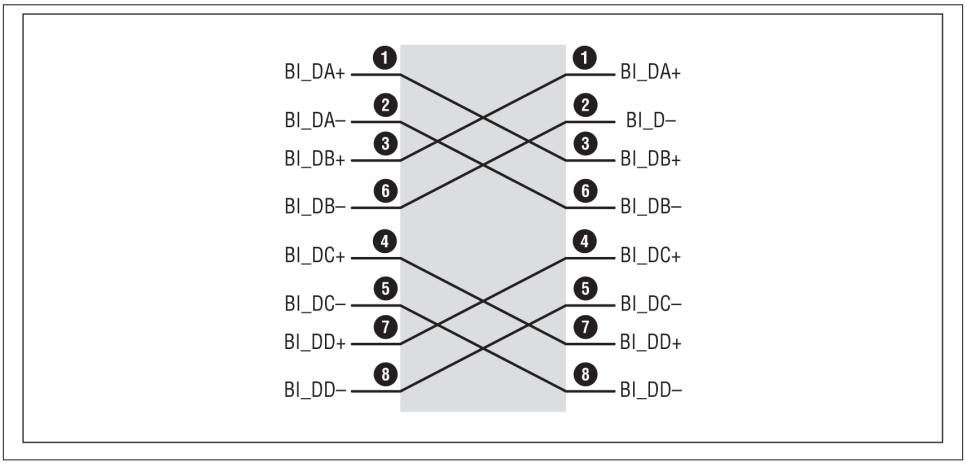


Figure 16-12. Four-pair crossover cable

Auto-Negotiation and MDIX Failures

You should know that disabling Auto-Negotiation can shut off Auto-MDIX under certain circumstances, which can cause failures on a link. This happens when you set a fixed speed by, for example, configuring a port to provide 10BASE-T operation only. The crossover failure occurs because disabling Auto-Negotiation removes support for 1000BASE-T on the switch port, which can *also* disable Auto-MDIX. The result is a link failure when the speed is manually configured, because the auto crossover mechanism that was making the link work is no longer available to automatically cross over the signals.

It can sometimes be difficult to tell how the Auto-Negotiation implementation on a given switch works from the documentation. It's therefore worthwhile to spend the time running some tests to see whether or not Auto-Negotiation and Auto-MDIX are still working after a manual speed or duplex setting is made on one end of the link.

Identifying a Crossover Cable

There are several ways to tell the difference between a normal straight-through cable and a crossover cable. Ideally, a crossover cable will be labeled as such at one or both ends of the cable, making identification easy. However, if there are no labels, then there are a couple of approaches you can take.

A handheld cable tester can be used to generate a "wiremap" of the cable, which typically provides a display that shows which wires are connected to which pins.

You can also try looking at the wire colors inside the RJ45-style plugs on each end of the cable, assuming that the plugs are made of transparent plastic. If you hold the two plugs together—side by side—you can see that the wire colors on the pins at each end of the cable are the same for a straight-through cable. On a crossover cable, the wire colors connected to pins 1 and 2 at one end of the cable will be connected to pins 3 and 6 at the other end.

CHAPTER 17

Fiber Optic Cables and Connectors

Fiber optic Ethernet cabling systems are based on multimode and single-mode fiber optic cables and connectors. Depending on the speed of the Ethernet media system and the type of fiber optic transmitters used in the system, the behavior of fiber optic media can vary. This is especially true for the high-speed Ethernet fiber optic systems, which send extremely rapid signals over fiber optic cable.

A major advantage of fiber optic cable is that the use of light pulses instead of electrical currents provides complete electrical isolation for equipment located at each end of a fiber optic link. This isolation provides immunity from hazards such as lightning strikes, and from the effects that can be caused by different levels of electrical ground potentials found in separate buildings.

For safe and reliable operation of your Ethernet system, electrical isolation of the sort provided by a fiber optic segment is essential when Ethernet segments are installed between buildings. Fiber optic media is also useful in environments such as manufacturing floors, because fiber optic segments are unaffected by the high levels of electrical noise that can be generated by heavy motors, welders, or other kinds of manufacturing equipment.

Fiber Optic Cable

There are a variety of fiber optic cable types; which you use is determined by the distance and speed required. The smallest cables are fiber optic patch cords, which often contain just 2 fibers, but may contain 12 or 24 fibers as needed on the 40 and 100 Gb/s Ethernet systems. Patch cords can be based on either multimode or single-mode fiber, depending on your requirements, and can also be equipped with whatever fiber optic connector you need at each end of the cable. Companies that build fiber optic cables can create cables to meet your requirements, and ship them to you within a day or two after you place the order.

Fiber optic horizontal cables for service to a work area are sometimes installed as part of a structured cabling system. However, in the vast majority of structured cabling systems today, Category 5e or 6A twisted-pair horizontal cables are used to deliver Ethernet signals to desktops and other devices.

Fiber optic backbone cables, on the other hand, are widely used in structured cabling systems to provide links between switches in telecommunications spaces within a building. These backbone cables usually contain 12 or 24 fibers, but they can contain more fibers as required. For large backbone cable installations, fiber optic cable manufacturers can build backbone cables to order, depending on your needs.

The ANSI/TIA-568 structured cabling standards, as described in Chapter 15, provide specifications for installing both backbone and horizontal segment fiber optic cables in a building.

Ethernet single-mode fiber optic equipment and other network devices based on single-mode fiber use laser light sources. Sufficiently powerful laser light can damage the retina of your eye without causing any feeling of pain, because the retina doesn't have any pain receptors.

Don't ever assume that it is safe to look into the end of a fiber optic cable. While Ethernet interfaces are designed to avoid sending full-power laser optical signals on disconnected interfaces, it's always possible that the fiber optic cable you are looking at is not connected to a device that throttles power.

Fiber optic Ethernet transceivers for use on multimode fiber are based on LED transmitters that emit a form of light that is not dangerous to the eye. However, you should treat all fiber optic cables with caution. Beware of looking directly into any fiber optic cable, and always observe safety precautions to protect your eyesight when working around fiber optic cable systems.

Fiber Optic Core Diameters

The thickness of the core optical fiber used in fiber optic cables is very small and is measured in millionths of a meter, called micrometers (μm), or microns. One type of multimode fiber optic cable that was popular in the past has a 62.5 μm fiber optic core and 125 μm outer cladding (62.5/125). Modern cabling systems are based on multimode cable with a 50 μm core and 125 μm cladding (50/125).

Single-mode fiber has a much smaller core with the same 125 μm thickness of outer cladding. Commercial single-mode fibers have core diameters that can vary from approximately 8–10 μm. They are collectively referred to as "10 micron fiber" in the stan-

dard. By way of comparison, a single sheet of copy paper has a thickness of roughly 100 µm.

Fiber Optic Modes

A fiber optic "mode" is a path that light can follow in traveling down a fiber. As the name implies, multimode cable has a larger core designed to support multiple modes, or paths, of light propagation.

When an incoherent light source such as an LED is coupled to multimode fiber, multiple paths of light from the LED are transmitted over the cable, as shown in Figure 17-1. An advantage of the larger core is easier coupling of the light source to the cable. A disadvantage is that the wider corridor for light transmission allows the multiple paths of LED light to bounce off the sides of the fiber.

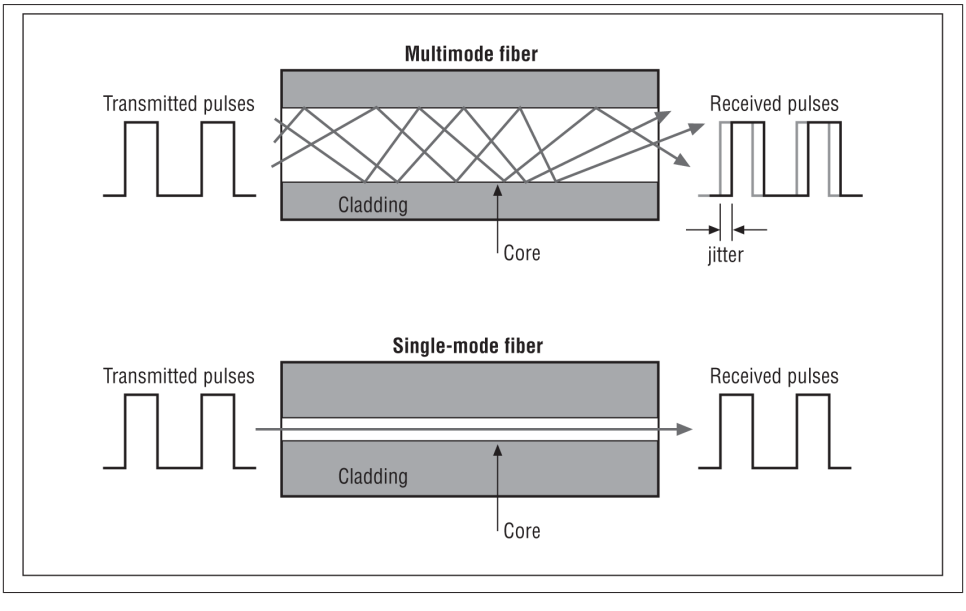


Figure 17-1. Modes of light

When this happens, these light paths arrive at the far end slightly out of phase, causing the light pulse to become dispersed, or spread out. This modal dispersion, or *jitter*, of the signal can cause problems with signal recovery at the far end. The longer the distance, the more signal dispersion there will be at a given signaling rate.

Single-mode fiber has a much smaller core, optimized to propagate a single mode, or path. When long-wavelength light (e.g., 1,300 nanometers) is injected into this fiber, only one mode will be active, and the rays of light will travel down the middle of the

fiber. When a coherent light source from a laser is coupled to a single-mode fiber, the single beam of laser light is transmitted over the single mode of the cable.

With single-mode fiber, the signals don't bounce against the cladding of the fiber, meaning there is no modal signal dispersion. Therefore, the light can travel a much longer distance without signal problems. The smaller core requires more precision to couple the light source to the cable, which is one reason why single-mode equipment is more expensive.

It is possible to couple a coherent laser light source to a multimode fiber. However, when this was first done for the Gigabit Ethernet media system, it was discovered that there can sometimes be a problem with signal propagation over the older cables that were common when Gigabit Ethernet was initally developed in 1999. This issue is called differential mode delay (DMD), and it is further described later in this chapter.

Since that time, newer versions of multimode cables have been developed that are designed to support laser light sources. In particular, these cables support the use of a less expensive laser technology called a Vertical Cavity Surface Emitting Laser (VCSEL). These lasers are well suited for low-cost transmission at the 850 nm wavelength, allowing for higher data rates over multimode fiber.

Fiber Optic Bandwidth

The distance over which an Ethernet signal will travel down a multimode fiber segment is primarily affected by signal strength and signal jitter, or dispersion. To help characterize the effects of dispersion, multimode fiber manufacturers specify their cables with a bandwidth rating, which is based on a figure of merit called the *bandwidth-distance product*, or simply *bandwidth*.

A figure of merit is a quantity used to characterize a component relative to its alternatives. Figures of merit are defined in engineering to provide a measure of a component's utility for the intended task. Examples include CPU speed, contrast ratios for LCD displays, and fiber optic dispersion ratings.

The bandwidth for multimode fiber optic cable is specified as the product of megahertz times kilometers, shown as either MHz-km or MHz*km. Single-mode cable does not have the modal dispersion issues of multimode fiber, and therefore is not provided with a bandwidth rating.

A 200 MHz-km fiber can move 200 MHz of data up to one kilometer or 100 MHz of data as far as two kilometers. The amount of modal dispersion is different at different frequencies of light; therefore, the bandwidth rating depends on the frequency of light

being sent over the cable. When using this spec, you need to know both the bandwidth rating and the frequency of light for which it applies on a given cable.

There is no way to field-test a fiber optic cable to derive a bandwidth-distance product. Instead, the bandwidth ratings can be found in vendor spec sheets (assuming you know the vendor and part number of the fiber cable). Multimode fiber optic media is manufactured in a range of bandwidths. A common rating for the older 62.5/125 µm cable was 160 MHz-km modal bandwidth at a wavelength of 850 nm. Newer versions of multimode fiber optic cables have since been developed with better ratings. To help keep things straight, the versions of multimode fiber have been provided with a rating system in the ISO/IEC 11801 standard based on the letters "OM," which stand for "optical multimode."

Table 17-1 shows the OM ratings for multimode fiber, showing the minimum modal bandwidth for two frequencies of operation.

Table 17-1. Optical specifications for multimode fiber

Fiber core diameter	ISO rating	MHz-km at 850 nm	MHz-km at 1,300 nm
62.5 µm MMF	OM1	200	500
50 µm MMF	OM2	500	500
50 µm MMF	OM3	1,500	500
50 µm MMF	OM4	3,500	500

Fiber manufacturers have improved the design and manufacturing of their cables since the older OM1 and OM2 cables were sold, including the development of laser-optimized multimode fiber (LOMF). The lower-cost light emitting diodes (LEDs) used in older OM1/OM2 fiber optic media systems for Ethernet have a maximum signaling rate of roughly 600 Mb/s. The VCSELs that are used to transmit signals over LOMF are capable of signaling at rates of over 10 Gb/s, which makes them ideal for use in high-speed networks.

Fiber Optic Loss Budget

The optical power losses on a fiber optic link must be small enough to allow the signal to be received accurately. The link power loss budget is the total optical power loss allocated for all fiber cables and patch cords and associated connectors on the segment, as well as all of the power penalties allocated in the standard to account for dynamic signal impairment factors. The power loss caused by the cable and connectors alone is called either *static power loss* or *channel insertion loss*.

The power penalties allocated for dynamic signal impairment cannot be measured in the field with static power loss testers. Instead, the dynamic power penalties are allocated in the standard to account for signaling issues such as modal noise, relative intensity noise, and intersymbol interference. The combined set of cable and connector losses

and the power penalties allocated for signaling losses make up the total optical power budget for the link.

The static loss, or channel insertion loss, includes the entire set of cables, patch cords, and connectors in the link. Optical power loss for a given type of fiber optic cable is expressed in dB/km (decibels per kilometer) at a specified wavelength. The "km" portion is assumed, and you will often see fiber optic loss measurements expressed using only the "dB" portion.

The channel insertion loss for a given link segment can be measured in the field with fiber optic test instruments that can tell you exactly how much optical loss there may be over a given segment at a given wavelength of light. The more connectors you have, and the longer your fiber link cable is, the higher the channel insertion loss will be. If the connectors or fiber splices are poorly made, or if there is finger oil and/or dust on the connector ends, then there will be higher optical loss on the segment.

When working with fiber optic cables, it is very important to keep the ends of the cables extremely clean. In addition, dust caps should be provided for any unused connectors to avoid any accumulation of dust and oil on the fiber optic equipment and cables. Fiber optic cleaning devices are available for cleaning the ends of fiber optic jumpers, cables, and transceiver ports before installation.

Note that fiber optic loss meters may use LED light sources operating at a typical wavelength of 850 nm. There is an issue when using such testers for faster Ethernet types, because they could produce an attenuation reading that would cause an otherwise acceptable link to be rejected. The faster Ethernet links (1 Gb/s and up) use laser light, which propagates more efficiently than LED light in most cases. Therefore, a loss reading performed with an LED-based tester can report a higher loss value than a tester with a laser light source.

Estimating the static optical loss

One way to provide a rough estimate of optical loss is to measure the segment length. Segment length is one of the most important cabling parameters for Ethernet fiber optic links, and the cable length will often determine whether a link will function (assuming that the optical losses of patch cables and connectors on a given segment are not excessive). Lower-cost field testers are available that can measure the length of an Ethernet segment, to help qualify it.

If the total segment length is acceptable, and if the link connectors and splices are correctly installed, then the link will probably work OK. If you have the test gear, then measuring the optical loss at the correct wavelength with the correct test device is the best method for determining optical loss. But in the absence of the correct test gear, optical cable length can provide a useful estimate.

If the link length is within the correct limits, but there is a high bit error rate or some other problem with the link, then to troubleshoot the issue you need to carefully check the optical attenuation. To get the most accurate attenuation test results, you must use a laser light source that operates at the same wavelength as the Ethernet media type that you intend to use.

Fiber Optic Connectors

A variety of fiber optic connectors are used, depending on the cable type and the Ethernet media system. The most commonly used fiber optic connectors as of this writing are the SC and LC connectors. You will also find ST connectors used on some older Ethernet equipment, and in older fiber optic cabling systems.

The SC connector is used on a variety of Ethernet transceivers. When higher density is needed to allow more ports within a given space, then the more compact LC connector is a popular choice. The Ethernet systems that operate at 40 and 100 Gb/s require multiple strands of fiber for their short reach media segments. The multiple-strand fiber cables, also known as "ribbon trunks," are terminated in multifiber push-on (MPO) connectors. An MPO connector can provide 12 or 24 fibers, depending on whether you are connecting to a 40 or 100 Gb/s short reach Ethernet interface.

ST Connectors

The ST ("straight tip") fiber optic connector is a registered trademark of AT&T Corp., formerly known as the American Telephone and Telegraph Company. The formal name of the ST connector in the ISO/IEC international standards is *BFOC/2.5*. ST connectors were once a popular choice for cabling system termination panels.

Figure 17-2 shows a pair of fiber optic cables equipped with ST plug connectors. The ST connector is a spring-loaded bayonet connector whose outer ring locks onto the connection. The ST connector has a key on an inner sleeve along with the outer bayonet ring.

To make a connection, you line up the key on the inner sleeve of the ST plug with a corresponding slot on the ST receptacle. Then you push the connector in and lock it in place by twisting the outer bayonet ring. This provides a tight connection with precise alignment between the two pieces of fiber optic cable being joined. The ST connector provides a very reliable connection that does not easily loosen or pop out of place.

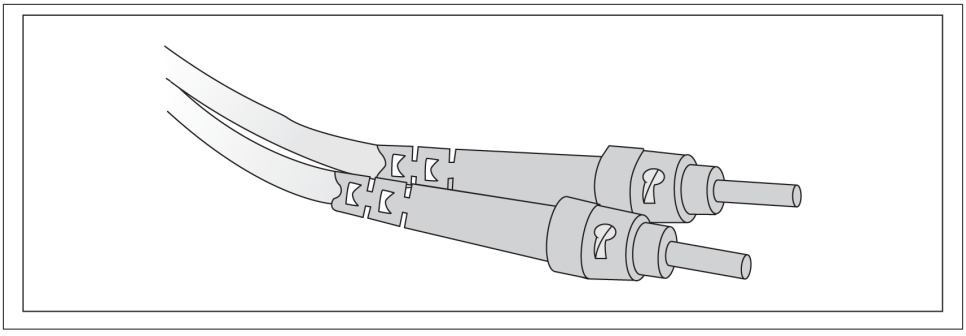


Figure 17-2. ST connectors

SC Connectors

SC, meaning "subscriber connector," is a registered trademark of Nippon Telegraph and Telephone (NTT). The SC connector is widely used by vendors for Ethernet interfaces, including long-range interfaces for 10, 40, and 100 Gb/s Ethernet.

The duplex SC connector shown in Figure 17-3 is also a recommended connector in the 100BASE-FX and 1000BASE-X standards.

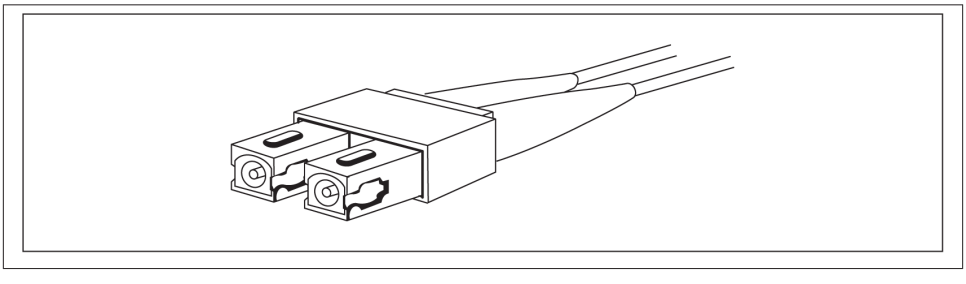


Figure 17-3. Duplex SC connector

The SC connector is designed for ease of use; the connector is pushed into place and automatically snaps into the connector housing to complete the connection. Make sure to seat the connector firmly, pushing until it has "clicked" into place. An SC connector might still work if it is not installed tightly, but you will encounter high error rates and eventually the link may fail completely.

LC Connectors

The LC connector was developed by Lucent, hence the name ("Lucent connector"). An LC connector uses a retaining tab mechanism, similar to an RJ45 connector, while the connector body has a square shape similar to the SC connector but smaller in size. LC

connectors are usually held together in a duplex configuration with a plastic clip, as shown in Figure 17-4. The ferrule of an LC connector is 1.25 mm.

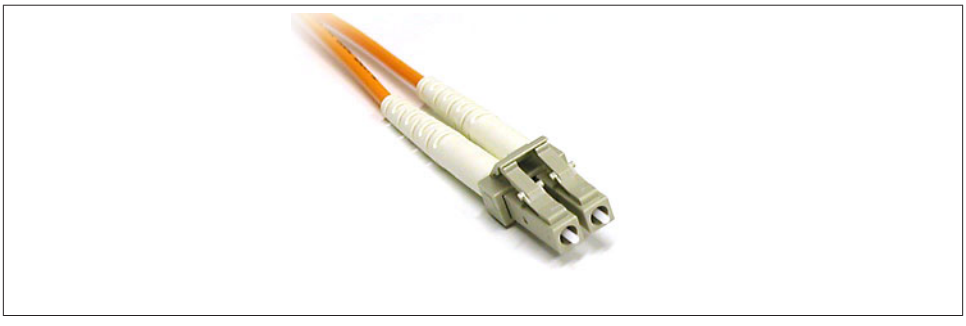


Figure 17-4. LC connector

The LC connector provides two fiber optic connections in a smaller space. Because the LC connector takes up about half the space required by an SC connector, this allows vendors to provide more ports on a switch front panel or chassis module.

MPO Connectors

As its name implies, the multifiber push-on (MPO) connector provides multiple fibers in a connector that is both "push to connect" and "push to disconnect." This connector is defined by IEC-61754-7, "Fibre optic interconnecting devices and passive components," and TIA-604-5-D, "Fiber Optic Connector Intermateability Standard, Type MPO." Both standards specify 12- and 24-fiber versions of the MPO connector.

You will also see the term MTP used for this connector type, which is a registered trademark of US Conec for a connector that is compliant with the MPO standards (meaning that the MTP connector is an MPO connector). However, the MTP connector has been enhanced by US Conec to provide several product features, including the ability to change gender or to repolish in the field, a floating ferrule for improved optical performance, and elliptical guide pins to provide for tighter tolerance in alignment. Some of these features are covered under patents.

Figure 17-5 shows two MPO connectors. One is a 12-fiber MPO plug, with the alignment pins sticking out. The other is a 24-fiber MPO jack, with alignment holes that mate with the alignment pins.

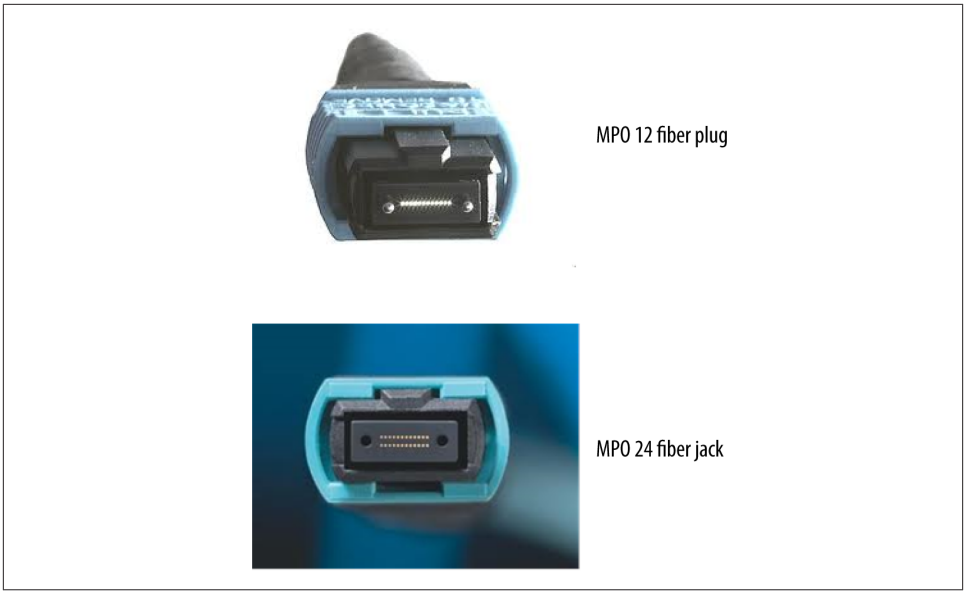


Figure 17-5. MPO connector

Building Fiber Optic Cables

Fiber optic patch cables can readily be purchased with fiber optic connectors already installed, allowing you to make relatively short-distance fiber connections quite easily. However, many fiber optic systems are installed to cover long distances between buildings or as backbone systems inside buildings. In this case, the typical installation is based on raw fiber, which is pulled into place and then terminated with fiber connectors that are installed in fiber optic patch panels. Long fiber optic segments may require the installation of several fiber optic cable segments, which are spliced together into a continuous cable.

There are a variety of fiber optic cable types and sizes, designed to meet virtually any installation requirement. While it's not rocket science, terminating a fiber optic cable in a connector and splicing raw fiber ends together requires specialized equipment and skills.

During installation, there are a number of special techniques that may be used for fiber optic cable splicing and terminating. Testing and verifying the operation of fiber optic cables also requires special equipment and training on how to operate the equipment. That's why the vast majority of sites turn to certified fiber optic installers and cable contractors for fiber optic installation and testing on segments that require cable termination and splicing.

Fiber Optic Color Codes

According to the TIA standard, unless the color coding is used for some other purpose, the connector plug body should be identified by the following colors where possible:

- Multimode: beige
- Single-mode: blue
- Angle Polished Contact (APC) single-mode connectors: green

The strain relief on the cable end and the cable jacket are also identified by a color code, as outlined in Table 17-2.

Table 17-2. Fiber jacket color code

Fiber type and class	Diameter	Jacket color
Multimode OM1/OM2	62.5/125 µm	Orange
Multimode OM2	50/125 µm	Orange
Laser-optimized Multimode OM3/OM4	50/125 µm	Aqua
Single-mode	10/125 µm	Yellow

Multifiber cables, also known as ribbon cables, have a color code for each of the fiber optic strands in the cable. As shown in Table 17-3, solid colors are used on the first 12 strands. If the ribbon cable has more than 12 strands, then the next 12 strands will have a solid color base and a "tracer" color, which is a second color that is marked with a dashed or continuous line, a dashed or spiral line, ring stripes, or hash marks.

Table 17-3. Multifiber color code

Position number	Color	Position number	Color and tracer
1	Blue	13	Blue with black tracer
2	Orange	14	Orange with black tracer
3	Green	15	Green with black tracer
4	Brown	16	Brown with black tracer
5	Slate	17	Slate with black tracer
6	White	18	White with black tracer
7	Red	19	Red with black tracer
8	Black	20	Black with yellow tracer
9	Yellow	21	Yellow with black tracer
10	Violet	22	Violet with black tracer
11	Rose	23	Rose with black tracer
12	Aqua	24	Aqua with black tracer

Signal Crossover in Fiber Optic Systems

A signal crossover is required to make a connection between an Ethernet transceiver attached to a station and the transceiver located in an Ethernet port. To make the data flow properly, the transmit data output of one transceiver must end up at the receive data input of the other transceiver. When connecting two nearby devices with a fiber optic patch cable, you must ensure that the transmit data pins at one device are connected to the receive data pins at the other device, and vice versa.

In the twisted-pair Ethernet system, signal crossover is done inside the Ethernet switch port using the Auto-MDIX system (as described in Chapter 16), and the structured cabling standard recommends wiring the twisted-pair cable segment straight through. Unlike with twisted-pair Ethernet, the structured cabling standard recommends that signal crossover for fiber optic horizontal segments be done in the cabling segment, and not at the Ethernet device.

For horizontal fiber optic segments installed as part of a structured cabling system, the connectors on the fiber optic cable should be oriented to achieve the crossover. For example, fiber optic cables are usually terminated in a set of fiber optic connectors located at the work area and the wiring closet. For a fiber optic cable with two optical fibers in it, fiber #1 is connected to the A connector at the work area end of the segment and the B connector at the wiring closet end. Fiber #2 is connected to the B connector at the work area, and the A connector at the wiring closet.

Using this method, the user or the network technician can connect fiber optic Ethernet ports to the fiber optic connectors on the horizontal cabling segment using straight-through fiber optic patch cables. They need not concern themselves about the signal crossover, because it is already accomplished in the fiber optic horizontal segment.

On the other hand, backbone fiber optic cable systems that go between floors of a building, or between buildings on a campus, are usually wired straight through. When making a connection between an Ethernet switch port and a backbone fiber optic system, you need to make sure that the signal crossover is achieved in the patch cable at one end of the link.

Signal Crossover in MPO Cables

Ribbon cables terminated with MPO connectors present some challenges when it comes to maintaining the correct signal crossover in a segment consisting of multiple fiber optic strands. The ANSI/TIA-568-C.3 standard provides a set of "Guidelines for Maintaining Polarity Using Array Connectors" that describe three types of MPO-to-MPO array cables, defined as Types A, B, and C. These cables are used to provide three different methods for maintaining a crossover connection. Method A is the preferred method, and is based on Type A MPO cables.

Figure 17-6 shows a Type A straight-through ribbon cable with 12 fibers terminated in MPO connectors. A Method A backbone link is cabled "straight through," terminating in the cabling system patch panel. One end of the link will have a straight-through patch cable, connecting from the patch panel to the Ethernet interface. The other end of the link will have a crossover cable connecting to the Ethernet interface. The guidelines recommend keeping all of the crossover patch cables at one end of the link, to keep the system as simple as possible and help the installer to avoid connecting the wrong type of patch cable.

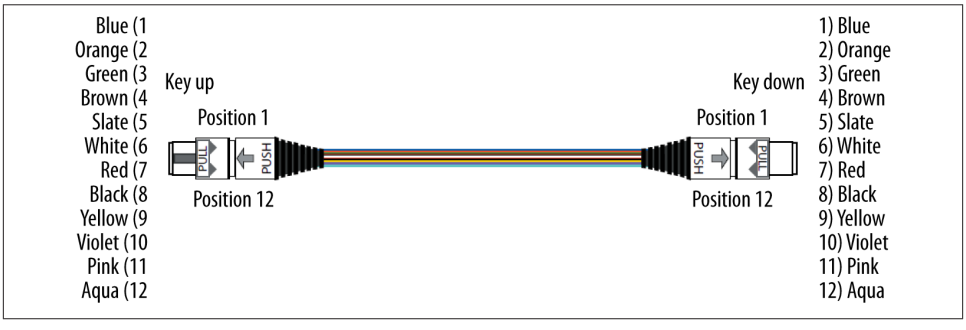


Figure 17-6. MPO Type A straight-through cable

The guidelines also show Method B and Method C, which are two methods for providing a crossover path built into the MPO backbone cables themselves. Given the complexity of these approaches and the difficulty of implementing them correctly, they are both rarely used.

As you can see, there are a variety of approaches to managing the signal crossover for the 12-fiber and 24-fiber systems needed to support 40 and 100 Gb/s Ethernet. For the best results, make sure you know which method your site is using in the cabling system, and order the correct MPO cable types to make the connections and achieve the signal crossover. Note that some vendors provide special MPO connectors that make it possible to change connector gender and polarity (crossover) in the field, which could be a handy way to resolve MPO-to-MPO connectivity issues.[1]

1. One example of this approach is available from Panduit.

PART IV
Ethernet Switches and Network Design

This section will describe the fundamentals of network design, and how to design and build Ethernet systems using Ethernet switches. Chapter 18 provides an introduction to the operation of Ethernet switches, while Chapter 19 covers the basics of network design with switches and provides brief descriptions of advanced switches.

CHAPTER 18
Ethernet Switches

Ethernet switches, also known as bridges, are basic building blocks of networks, and are so commonly used that you may not give them a second thought. It's possible to build networks without knowing very much about how switches work. However, when you build larger network systems, it helps to understand both what goes on inside a switch and how the standards make it possible for switches to work together.

Ethernet is used to build networks from the smallest to the largest, and from the simplest to the most complex. Ethernet connects your home computers and other household devices; switches for home networks are typically small, low-cost, and simple. Ethernet also connects the worldwide Internet, and switches for Internet service providers are large, high-cost, and complex.

Campus and enterprise networks often use a mix of switches, with simpler and lower-cost switches used inside wiring closets to connect devices on a given floor of a building, and larger and higher-cost switches in the core of the network to connect all the building switches together into a larger network system. Data center networks have their own special requirements, and typically include high-performance switches that can be connected in ways that provide highly resilient networks.

According to industry estimates, the worldwide market for enterprise switches had revenues of over $5 billion per quarter in 2013, with total revenues exceeding $20 billion for the year. For the third quarter of 2013, there were tens of millions of Ethernet ports shipped, including 4.7 million 10 Gigabit ports. To satisfy the large and ever-increasing market for Ethernet switches, there are many varieties of switches offered at many price points.

The many kinds of switches, and the many features that can be found in those switches, is a big topic. Covering the entire range of technology and the various ways switches can be used in network designs would require an entire book, or even several books. Instead, in this chapter we will provide an introduction to and a brief tutorial on how

switches function. In Chapter 19, we will provide a tutorial on using switches in network designs and an overview of the most useful features for network design that are found in switches, including the basic features included in most switches and the more advanced features found in higher-cost and specialized switches.

Basic Switch Functions

Ethernet switches link Ethernet devices together by relaying Ethernet frames between the devices connected to the switches. By moving Ethernet frames between the switch ports, a switch links the traffic carried by the individual network connections into a larger Ethernet network.

Ethernet switches perform their linking function by *bridging* Ethernet frames between Ethernet segments. To do this, they copy Ethernet frames from one switch port to another, based on the media access control (MAC) addresses in the Ethernet frames. Ethernet bridging was initially defined in IEEE Standard 802.1D, "IEEE Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges."

The most recent version of the 802.1D bridging standard is dated 2004. The 802.1D standard was extended and enhanced by the subsequent development of the 802.1Q-2011 standard, "Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks."

The standardization of bridging operations in switches makes it possible to buy switches from different vendors that will work together when combined in a network design. That's the result of lots of hard work on the part of the standards engineers to define a set of standards that the vendors can agree upon and implement in their switch designs.

Bridges and Switches

The first Ethernet bridges were two-port devices that could link two of the original Ethernet system's coaxial cable segments together. At that time, Ethernet only supported connections to coaxial cables. Later, when twisted-pair Ethernet was developed and switches with many ports became widely available, they were often used as the central connection point, or hub, of Ethernet cabling systems, resulting in the name "switching hub." Nowadays, in the marketplace, these devices are simply called switches.

Things have changed quite a lot since Ethernet bridges were first developed in the early 1980s. Over the years, computers have become ubiquitous, and many people use multiple devices at their jobs, including laptops, smartphones, and tablets. Every VoIP telephone and every printer is a computer, and even building management systems and access controls (door locks) are networked. Modern buildings have multiple wireless access points (APs) to provide 802.11 Wi-Fi services for devices like smartphones and

tablets, and each of the APs is also connected to a cabled Ethernet system. As a result, modern Ethernet networks may consist of hundreds of switch connections in a building, and thousands of switch connections across a campus network.

What Is a Switch?

You should know that there is another network device used to link networks, called a *router*. There are major differences in the way that bridges and routers work, and they both have advantages and disadvantages, as you'll see. Very briefly, bridges move frames between Ethernet segments based on Ethernet addresses, with little or no configuration of the bridge required. Routers move packets between networks based on high-level protocol addresses, and each network being linked must be configured into the router. However, both bridges and routers are used to build larger networks, and both devices are called switches in the marketplace.

We will use the words "bridge" and "switch" interchangeably to describe Ethernet bridges. However, note that "switch" is a generic name for network devices that may function as bridges, routers, or both, depending on their feature set and configuration. The point is that as far as network experts are concerned, bridging and routing are different kinds of packet switching with different capabilities. For our purposes, we will follow the practices of Ethernet vendors who use the word "switch," or more specifically, "Ethernet switch," to describe devices that bridge Ethernet frames.

While the 802.1D standard provides the specifications for bridging local area network (LAN) frames between ports of a switch, and for a few other aspects of basic bridge operation, the standard is also careful to avoid specifying issues like bridge or switch performance or how switches should be built. Instead, vendors compete with one another to provide switches at multiple price points and with multiple levels of performance and capabilities.

The result has been a large and competitive market in Ethernet switches, increasing the number of choices you have as a customer. The wide range of switch models and capabilities can be confusing, so in Chapter 19, we will review the different kinds of switches.

Operation of Ethernet Switches

Networks exist to move data between computers, and to perform that task, the data being moved is organized as chunks of data called Ethernet *frames*. Frames travel over Ethernet networks, and the data field of a frame is used to carry data between computers. Frames are nothing more than an arbitrary sequence of information whose format is defined in a standard.

As shown in Figure 18-1, the format for an Ethernet frame includes a destination address as the first field received, containing the address of the device to which the frame is being sent. (The preamble field at the beginning of the frame is automatically stripped

off when the frame is received on an Ethernet interface, leaving the destination address as the first field.)

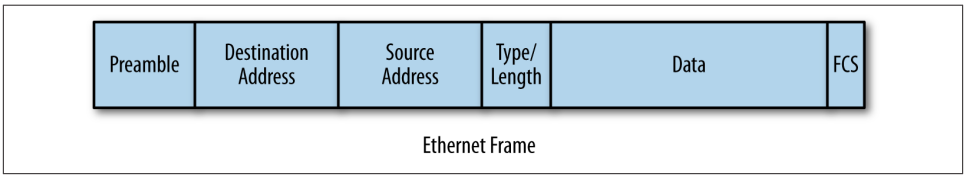


Figure 18-1. Ethernet frame format

Next will be a source address, containing the address of the device sending the frame. The addresses will be followed by various other fields, including the data field, which carries the data being sent between computers. (For a full discussion of the structure of an Ethernet frame, refer to Chapter 4.)

Frames are defined at Layer 2, the data link layer, of the OSI seven-layer network model. As discussed in Chapter 2, the seven-layer model was developed to organize the kinds of information sent between computers, which helps to define how that information will be sent and to structure the development of standards for that task. Because Ethernet switches operate on local area network frames at the data link layer so you will sometimes hear them called "link layer devices," as well as "Layer 2 devices" or "Layer 2 switches."

In essence, Ethernet functions as the trucking system that transports TCP/IP packets between computers, carried as data in the Ethernet frames. Although you will also hear Ethernet frames referred to as "packets," as far as the standards are concerned, Ethernet uses frames to carry data between computers.

The TCP/IP network protocol is based on network-layer *packets*. The TCP/IP packets are carried between computers in the data fields of Ethernet *frames*.

Ethernet switches are designed so that their operation is invisible to the devices on the network, which explains why this approach to linking networks is also called *transparent bridging*. "Transparent" means that when you connect a switch to an Ethernet system, no changes are made in the Ethernet frames that are bridged. The switch will automatically begin working; no configuration on the switch is required and you don't need to make any changes on the computers connected to the Ethernet network, making the operation of the switch transparent to them.

Next, we will look at the basic functions used in a bridge to make it possible to forward Ethernet frames from one port to another.

Address Learning

An Ethernet switch controls the transmission of frames between switch ports connected to Ethernet cables using the traffic forwarding rules described in the IEEE 802.1D bridging standard. Traffic forwarding is based on address learning. Switches make traffic forwarding decisions based on the 48-bit MAC addresses used in LAN standards including Ethernet.

To do this, the switch learns which devices, called "stations" in the standard, are on which segments of the network by looking at the source addresses in all of the frames it receives. When an Ethernet device sends a frame, it puts two addresses in the frame. These two addresses are the *destination* address of the device it is sending the frame to, and the *source* address, which is the address of the device sending the frame.

The way the switch "learns" is fairly simple. Like all Ethernet interfaces, every port on a switch has a unique factory-assigned MAC address. However, unlike a normal Ethernet device that accepts frames addressed directly to it, the Ethernet interface located in each port of a switch runs in *promiscuous* mode. In this mode, the interface is programmed to receive *all* frames it sees on that port, not just the frames that are being sent to the MAC address of the Ethernet interface on that switch port.

As each frame is received on each port, the switching software looks at the source address of the frame and adds that source address to a table of addresses that the switch maintains. This is how the switch automatically discovers which stations are reachable on which ports.

Figure 18-2 shows a switch linking six Ethernet devices. For convenience, we're using short numbers for station addresses, instead of actual 6-byte MAC addresses. As stations send traffic, the switch receives every frame sent and builds a table, more formally called a *forwarding database*, that shows which stations can be reached on which ports.

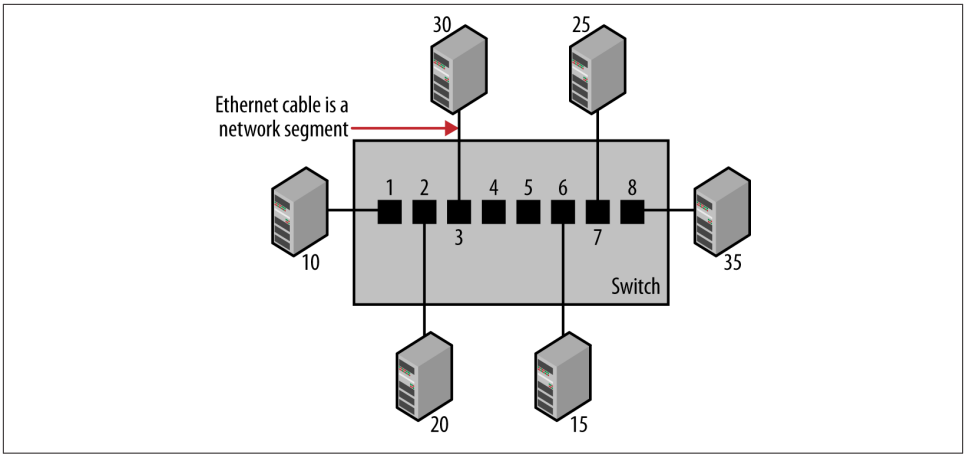


Figure 18-2. Address learning in a switch

After every station has transmitted at least one frame, the switch will end up with a forwarding database such as that shown in Table 18-1.

Table 18-1. Forwarding database maintained by a switch

Port	Station
1	10
2	20
3	30
4	No station
5	No station
6	15
7	25
8	35

The switch then uses this database to make packet forwarding decisions, in a process called *adaptive filtering*. Without an address database, the switch would have to send traffic received on any given port out all other ports to ensure that it reached its destination. With the address database, the traffic is filtered according to its destination. The switch is "adaptive" in that it learns new addresses automatically.

This ability to learn makes it possible for you to add new stations to your network without having to manually configure the switch to know about the new stations, or for the stations to know about the switch.

Any Ethernet system still using coaxial cable segments and/or repeater hubs may have multiple stations on a network segment. Connecting such a segment to a switch will result in multiple stations being reachable over a single port.

When the switch receives a frame that is destined for a station address that it hasn't yet seen, the switch will send the frame out all of the ports other than the port on which it arrived. This process is called flooding, and is explained in more detail later. Suppressing frame transmission on the switch port that receives the frame prevents stations on a shared segment connected to that port from seeing the same traffic more than once. This also prevents a single station on a port from receiving a copy of the frame it just sent.

Traffic Filtering

Once the switch has built a database of addresses, it has all the information it needs to filter and forward traffic selectively. While the switch is learning addresses, it is also checking each frame to make a packet forwarding decision based on the destination address in the frame. Let's look at how the forwarding decision works in a switch equipped with eight ports, like the one shown in Figure 18-2.

Assume that a frame is sent from station 15 to station 20. Because the frame is sent by station 15, the switch reads the frame in on port 6 and uses its address database to determine which of its ports is associated with the destination address in this frame. Here, the destination address corresponds to station 20, and the address database (Table 18-1) shows that to reach station 20, the frame must be sent out port 2.

Each port in the switch is provided with the ability to hold a small amount of data in memory, which holds the frame before transmission onto the Ethernet cable connected to the port. If the port is already busy transmitting when a frame arrives for transmission, then the frame can be held for the short time it takes for the port to complete transmitting the previous frame. To transmit the frame, the switch places the frame into the *packet switching queue* for transmission on port 2.

During this process, a switch transmitting an Ethernet frame from one port to another makes no changes in the data, address, or other fields of the basic Ethernet frame. Using our example, the frame is transmitted intact on port 2 exactly as it was received on port 6. Therefore, the operation of the switch is transparent to all stations on the network.

Note that the switch will not forward a frame destined for a station that is in the forwarding database onto a port unless that port is connected to the target destination. In other words, traffic destined for a device on a given port will only be sent to that port; no other ports will see the traffic intended for that device. This switching logic keeps

traffic isolated to only those Ethernet cables, or segments, needed to receive the frame from the sender and transmit that frame to the destination device.

This prevents the flow of unnecessary traffic on other segments of the network system, which is a major advantage of a switch. This is in contrast to the early Ethernet systems, where traffic from any station was seen by all other stations, whether they wanted it or not. Switch traffic filtering reduces the traffic load carried by the set of Ethernet cables connected to the switch, thereby making more efficient use of the network bandwidth.

Frame Flooding

Switches automatically age out entries in their forwarding databases after a period of time, usually five minutes, if they do not see any frame from a station. If a station doesn't send traffic for a designated period, then the switch will delete the forwarding entry for that station. This keeps the forwarding database from growing full of stale entries that might not reflect reality.

However, once an address entry has timed out the next time the switch receives a frame destined for the station whose entry has timed out, it won't have any information for that station in the database. This also happens when a station is newly connected to a switch, or when a station has been powered off and is turned back on more than five minutes later. So how does the switch handle packet forwarding for an unknown station?

The solution is simple: the switch forwards the frame destined for an unknown station out all switch ports other than the one it was received on, *flooding* the frame to all other stations. Flooding the frame guarantees that a frame with an unknown destination address will reach all network connections and be heard by the correct destination device, assuming that it is active and on the network. When the unknown device responds with return traffic, the switch will automatically learn which port the device is on, and will no longer flood traffic destined for that device.

Broadcast and Multicast Traffic

In addition to frames directed to a single address, local area networks are capable of sending frames directed to a group address, called a *multicast* address, which can be received by a group of stations, as well as frames directed to all stations, called the *broadcast* address. Group addresses always begin with a specific bit pattern defined in the Ethernet standard, making it possible for a switch to determine which frames are destined for a specific device versus a group of devices.

A frame sent to a multicast destination address can be received by all stations configured to listen for that multicast address. The Ethernet software, also called "interface driver" software, programs the interface to accept frames sent to the group address, so that the interface is now a member of that group. The Ethernet interface address assigned at the factory is called a *unicast* address, and any given Ethernet interface can receive unicast

frames and multicast frames. In other words, the interface can be programmed to receive frames sent to one or more multicast group addresses, as well as frames sent to the unicast MAC address belonging to that interface.

Broadcast and multicast forwarding

The broadcast address is the group of all stations, which is a special case of multicast. A packet sent to the broadcast address (the address of all ones) is received by every station on the LAN. Because broadcast packets must be received by all stations on the network, the switch will achieve that goal by flooding broadcast packets out all ports except the port that it was received on—there's no need to send the packet back to the originating device. This way, a broadcast packet sent by any station will reach all other stations on the LAN.

Multicast traffic can be more difficult to deal with than broadcast frames. More sophisticated (and usually more expensive) switches include support for multicast group discovery protocols that make it possible for the station to tell the switch about the multicast group addresses that the station wants to receive frames from, so that the switch sends multicast packets only to the ports connected to stations that have indicated their interest in receiving the multicast traffic. However, lower-cost switches with no capability to discover which ports are connected to stations listening to a given multicast address must resort to flooding multicast packets out all ports other than the port on which the multicast traffic was received, just like broadcast packets.

Uses of broadcast and multicast

Stations send broadcast and multicast packets for a number of reasons. High-level network protocols like TCP/IP use broadcast or multicast frames as part of their address discovery process. Broadcasts and multicasts are also used for dynamic address assignment on stations, which occurs when a station is first powered on and needs to find a high-level network address. Multicasts are also used by certain multimedia applications, which send audio and video data in multicast frames for reception by groups of stations, and by multiuser games as a way of sending data to the group of game players.

Therefore, a typical network will have some level of broadcast and multicast traffic. As long as the number of such frames is at a reasonable level, then there won't be any problems. However, when many stations are combined by switches into a single large network, broadcast and multicast flooding by the switches can result in significant amounts of traffic. Large amounts of broadcast or multicast traffic may cause network congestion, because every device on the network is required to receive and process broadcasts and specific types of multicasts; at high enough packet rates, there could be performance issues on the stations.

Streaming applications (video) sending high rates of multicasts can generate intense traffic. Disk backup and disk duplication systems based on multicast can also generate

lots of traffic. If this traffic ends up being flooded to all ports, the network could get congested. One way to avoid this congestion is to limit the total number of stations linked to a single network, so that the broadcast and multicast rate does not get so high as to be a problem.

Another way to limit the rate of multicast and broadcast packets is to divide the network into multiple virtual LANs (VLANs), each of which operates as a separate and distinct LAN. Yet another method is to use a router, also called a Layer 3 switch. Because a router does not automatically forward broadcasts and multicasts between networks, this creates separate broadcast domains. Both methods for controlling the propagation of multicasts and broadcasts are discussed in more detail later (VLANs in this chapter and routers in the next).

Combining Switches

So far we've seen how a single switch can forward traffic based on a dynamically created forwarding database. A major difficulty with this simple model of switch operation is that multiple connections between switches can create loop paths, leading to network congestion and overload.

Forwarding Loops

The design and operation of Ethernet requires that only a single packet transmission path may exist between any two stations. An Ethernet grows by extending branches in a network topology called a *tree structure*, which consists of multiple switches branching off of a central switch. The danger is that in a sufficiently complex network, switches with multiple interswitch connections can create loop paths in the network.

On a network with switches connected together so that a packet forwarding loop is formed, packets will circulate endlessly around the loop, building up to very high levels of traffic and causing an overload.

The looped packets will circulate at the maximum rate of the network links, until the traffic rate gets so high that the network is saturated. Broadcast and multicast frames are normally flooded to all ports in a basic switch, as are unicast frames being sent to unknown destinations, and all of this traffic will circulate in such a loop. Once a loop is formed, this failure mode can happen very rapidly: the network will be fully occupied sending broadcast, multicast, and unknown frames, making it very difficult for stations to send unicast traffic destined for known stations.

Unfortunately, loops like the dotted path shown with arrows in Figure 18-3 are all too easy to achieve, despite your best efforts to avoid them. As networks grow to include more switches and more wiring closets, it becomes difficult to know exactly how things are connected together and to keep people from mistakenly creating loop paths.

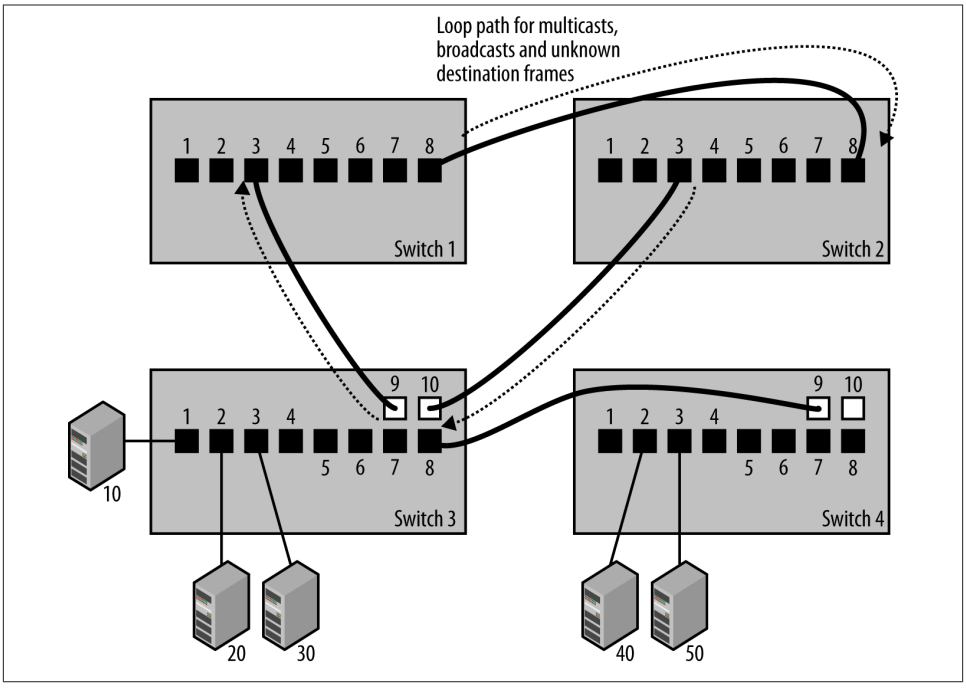


Figure 18-3. Forwarding loop between switches

While the loop in the drawing is intended to be obvious, in a sufficiently complex network system it can be challenging for anyone working on the network to know whether or not the switches are connected in such a way as to create loop paths. The IEEE 802.1D bridging standard provides a Spanning Tree Protocol (STP) to avoid this problem by automatically suppressing forwarding loops; we'll explore this protocol next.

The Spanning Tree Protocol

The purpose of the Spanning Tree Protocol is to allow switches to automatically create a loop-free set of paths, even in a complex network with multiple paths through multiple switches. The ability to dynamically create a tree topology in a network by blocking any packet forwarding on certain ports is the mechanism that ensures that a set of Ethernet switches can automatically configure themselves to produce loop-free paths. The IEEE 802.1D standard describes the operation of spanning tree, and every switch that claims compliance with the 802.1D standard must including spanning tree capability.

Beware that low-cost switches may not include spanning tree capability, rendering them unable to block any packet forwarding loops. Also, some vendors that provide this capability may disable it by default, requiring you to manually enable the Spanning Tree Protocol before it will function to protect your network.

Spanning tree packets

Operation of the spanning tree algorithm is based on configuration messages sent by each switch in packets called *bridge protocol data units* (BPDUs). Each BPDU packet is sent to a destination multicast address that has been assigned to spanning tree operation. All IEEE 802.1D switches join the BPDU multicast group and listen to frames sent to this address, so that every switch can send and receive spanning tree configuration messages. Figure 18-4 illustrates how this works.

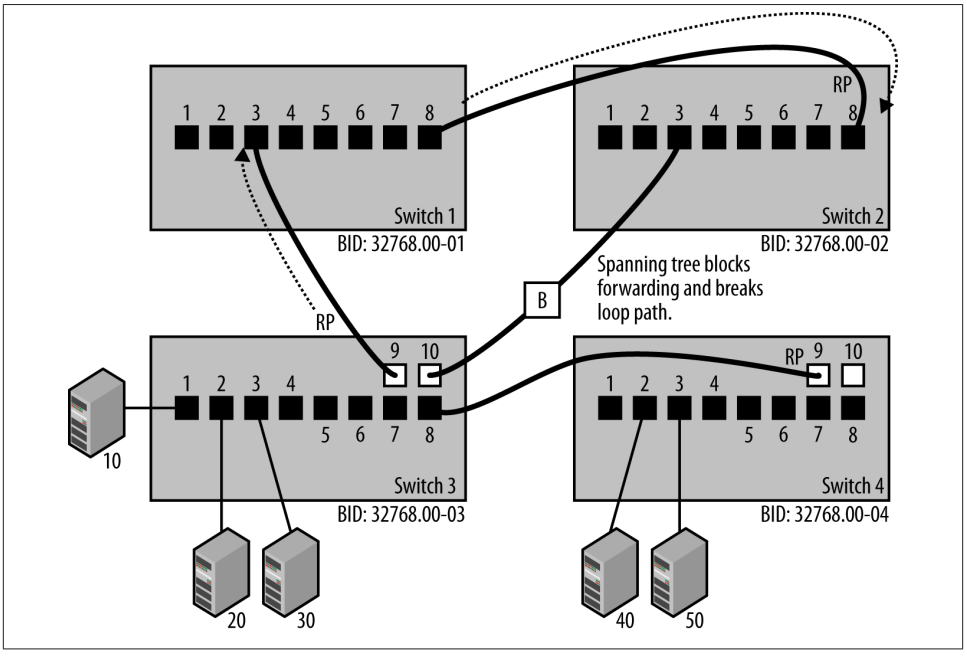


Figure 18-4. Spanning tree operation

The bridge multicast group MAC address is 01-80-C2-00-00-00. Vendor-specific spanning tree enhancements may also use other addresses. For example, Cisco's Per-VLAN Spanning Tree (PVST) sends BPDUs to address 01-00-0C-CC-CC-CD.

Choosing a root bridge

The process of creating a spanning tree begins by using the information in the BPDU configuration messages to automatically elect a *root bridge*. The election is based on a bridge ID (BID) that, in turn, is based on the combination of a configurable bridge priority value (32,768 by default) and the unique Ethernet MAC address, called the *system MAC*, assigned on each bridge for use by the spanning tree process. Bridges send BPDUs to one another, and the bridge with the lowest BID is automatically elected to be the root bridge.

Assuming that the bridge priority was left at the default value of 32,768, then the bridge with the lowest numerical value Ethernet address will be the one elected as the root bridge.

It may happen that a low-performance bridge on your network will have the lowest MAC address and end up as the root bridge. You can configure a lower bridge priority on your core bridge to ensure that the core bridge is chosen to be root and that the root will be located at the core of your network and running on the higher-performance switch located there.

In the example shown in Figure 18-4, Switch 1 has the lowest BID, and the end result of the spanning tree election process is that Switch 1 has become the root bridge. Electing the root bridge sets the stage for the rest of the operations performed by the Spanning Tree Protocol.

Choosing the least-cost path

Once a root bridge is chosen, each non-root bridge uses that information to determine which of its ports has the least-cost path to the root bridge, and assigns that port to be the root port (RP). All bridges also compute the lowest-cost path to the root bridge for their other connections, and mark each as a designated port (DP). The bridge with the DP is the designated bridge (DB).

The *path cost* is based on the speed at which the ports operate, with higher speeds resulting in lower cost. As BPDU packets travel through the system, they accumulate information about the number of ports they travel through and the speed of each port. Paths with slower-speed ports will have higher costs. The total cost of a given path through multiple switches is the sum of the costs on all network segments on that path. If there are multiple paths to the root with the same cost, then the path connected to the bridge with the lowest bridge ID will be used.

At the end of this process, the bridges have chosen a set of root ports and designated ports, making it possible for the bridges to remove all loop paths and maintain a packet

forwarding tree that spans the entire set of devices connected to the network—hence the name Spanning Tree Protocol.

Blocking loop paths

Once the spanning tree process has determined the port status, the combination of root ports and designated ports provides the spanning tree algorithm with the information it needs to identify the best paths. Packet forwarding on any port that is not a root port or a designated port is disabled by *blocking* the forwarding of packets on that port.

While blocked ports do not forward packets, they continue to receive BPDUs. The blocked port is shown in Figure 18-4 with a "B," indicating that port 10 on Switch 3 is in blocking mode and that the link is not forwarding packets. The Rapid Spanning Tree Protocol (RSTP) sends BPDU packets every two seconds to monitor the state of the network, and a blocked port may become unblocked when a path change is detected.

Spanning tree port states

When an active device is connected to a switch port, the port goes through a number of states as it processes any BPDUs that it receives and the spanning tree process determines what state the port should be in. Figure 18-5 shows the various port states. During the *listening* and *learning* states, the spanning tree process on that port listens for BPDUs and learns source addresses from any frames received.

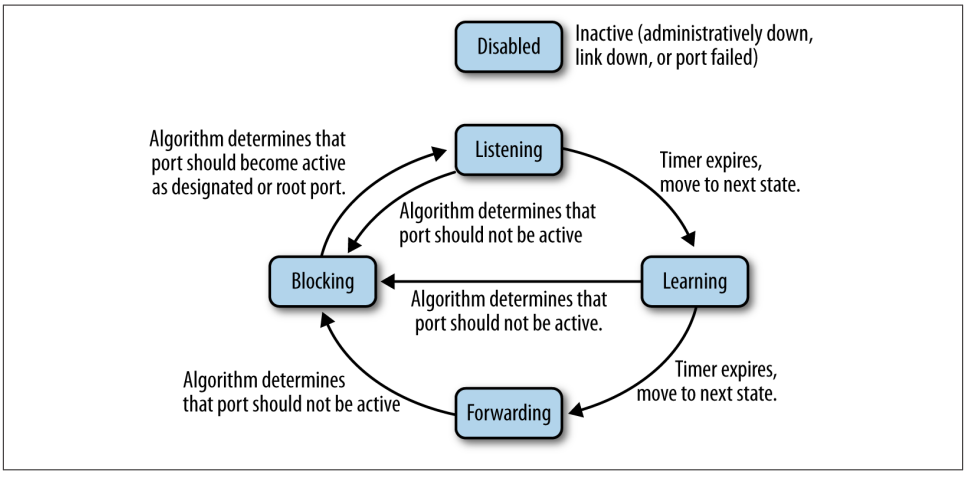


Figure 18-5. Spanning tree port states

The state diagram shows the spanning tree port states, which include the following:

Disabled
: A port in this state has been intentionally shut down by an administrator, or has automatically shut down because the link was disconnected or because the port has failed and is no longer operational. The disabled state can be entered from or left for any other state.

Blocking
: A port that is enabled but is not a root port or designated port could cause a switching loop if it were active (i.e., in the forwarding state). To avoid that, the port is placed in the blocking state. No station data is sent or received over a blocking port. Upon initialization of a port (link comes up, power is turned on), the port will typically enter the blocking state. Upon discovering via BPDUs or timeouts that the port may need to become active, the port will move to the listening state on the way to the forwarding state. A blocking port may also transition to the forwarding state if other links fail. BPDU data is still received while a port is in the blocking state.

Listening
: In this state, the port discards traffic but continues to process any BPDUs it receives and acts on any new information that would cause the port to return to the blocked state. Based on information received in BPDUs, the port may transition to the learning state. The listening state allows the spanning tree algorithm to decide whether the attributes of this port, such as port cost, should cause the port to become part of the spanning tree or return to blocking state.

Learning
: In this state, the port does not yet forward frames, but it does learn source addresses from any frames received and adds them to the filtering database. The switch will populate the MAC address table with packets heard on the port until the timer expires before moving to the forwarding state.

Forwarding
: This is the operational state in which a port is sending and receiving station data. Incoming BPDUs are also monitored to allow the spanning tree process on the bridge to detect if it needs to move the port into the blocking state to prevent a loop.

In the original Spanning Tree Protocol, the listening and learning states lasted for 30 seconds, during which time packets were not forwarded. In the newer Rapid Spanning Tree Protocol, it is possible to assign a port type of "edge" to a port, meaning that the port is known to be connected to an end station (user computer, VoIP telephone, printer, etc.) and not to another switch. That allows the RSTP state machine to bypass the learning and listening process on that port and to transition it to the forwarding state immediately.

Allowing a station to immediately begin sending and receiving packets helps avoid such issues as application timeouts on user computers when they are rebooted.

Prior to the development of RSTP, some vendors had developed their own versions of this feature. Cisco Systems, for example, provided the "portfast" command to enable an edge port to immediately begin forwarding packets.

While not required for RSTP operation, it is useful to manually configure RSTP edge ports with their port type, to avoid issues on user computers. Setting the port type to edge also means that the RSTP state machine doesn't need to send a BPDU packet upon link state change (link up or down) on that port, which helps reduce the amount of spanning tree traffic in your network.

The inventor of the Spanning Tree Protocol, Radia Perlman, wrote a poem to describe how it works.[1] When reading the poem it helps to know that in math terms, a network can be represented as a type of graph called a mesh, and that the goal of the Spanning Tree Protocol is to turn any given network mesh into a tree structure with no loops that spans the entire set of network segments.

I think that I shall never see
A graph more lovely than a tree.
A tree whose crucial property
Is loop-free connectivity.
A tree that must be sure to span
So packets can reach every LAN.
First, the root must be selected.
By ID, it is elected.
Least cost paths from root are traced.
In the tree, these paths are placed.
A mesh is made by folks like me,
Then bridges find a spanning tree.

—Radia Perlman
Algorhyme

This brief description is only intended to provide the basic concepts behind the operation of the system. As you might expect, there are more details and complexities that are not described here. The complete details of how the spanning tree state machine operates are described in the IEEE 802.1 standards, which can be consulted for a more complete understanding of the protocol and how it functions. The details of vendor-

1. Radia Perlman. *Interconnections: Bridges, Routers, Switches, and Internetworking Protocols, 2nd ed.*, (New York: Addison-Wesley, 1999), 46.

specific spanning tree enhancements can be found in the vendor documentation. See Appendix A for links to further information.

Spanning Tree Versions

The original Spanning Tree Protocol, standardized in IEEE 802.1D, specified a single spanning tree process running on a switch, managing all ports and VLANs on the switch with a single spanning tree state machine. Nothing in the standard prohibits vendors from developing their own enhancements to how spanning tree is deployed, and some have created their own implementations, in one case providing a separate spanning tree process per VLAN. That approach was taken by Cisco Systems for a version it calls Per-VLAN Spanning Tree (PVST). A "VLAN-aware" spanning tree protocol such as PVST can take VLANs into account when managing redundant paths.

The IEEE standard Spanning Tree Protocol has evolved over the years. An updated version, called the Rapid Spanning Tree Protocol, was defined in 2004. As the name implies, RSTP has increased the speed at which the protocol operates. The 802.1Q standard includes both RSTP and a new version of spanning tree called multiple spanning tree (MST), both of which were designed to be backward compatible with previous versions of the protocol.[2] MST is discussed further in "802.1Q Multiple Spanning Tree Protocol" on page 325.

When building a network with multiple switches, you need to pay careful attention to how the vendor of your switches has deployed spanning tree, and to the version of spanning tree your switches may have. The most commonly used versions, classic STP and the newer RSTP, are interoperable and both require no configuration, resulting in "plug and play" operation.

Before putting a new switch into operation on your network, read the vendor's documentation carefully and make sure that you understand how things work. Some vendors may not enable spanning tree on all ports by default. Other vendors may implement special features or vendor-specific versions of spanning tree.

Typically, a vendor will work hard to make sure that its implementation of spanning tree "just works" with all other switches, but there are enough variations in spanning tree features and configuration that you may encounter issues. Reading the documentation and testing new switches before deploying them throughout your network can help avoid any problems.

2. The IEEE 802.1Q standard (Note 1, p. 319) notes that "The spanning tree protocols specified by this standard supersede the Spanning Tree Protocol (STP) specified in IEEE Std 802.1D revisions prior to 2004, but facilitate migration by interoperating with the latter."

Switch Performance Issues

A single full-duplex Ethernet connection is designed to move Ethernet frames between the Ethernet interfaces at each end of the connection. It operates at a known bit rate and a known maximum frame rate.[3] All Ethernet connections at a given speed will have the same bit rate and frame rate characteristics. However, adding switches to the network system creates a more complex system. Now, the performance limits of your network are determined by the performance of the Ethernet connections and the performance of the switches, as well as any congestion that may occur in the system, depending on the topology. It's up to you to make sure that the switches you buy are capable of doing the job.

The internal switching electronics may not be able to sustain the full frame rate coming in from all ports. In other words, should all ports simultaneously present high traffic loads to the switch that are continual and not just short bursts, the switch may not be able to handle the combined traffic rate and may begin dropping frames. This is known as *blocking*, a condition that occurs in a switching system when there are insufficient resources available to provide for the flow of data through the switch. A *non-blocking switch* is one that provides enough internal switching capability to handle the full load even when all ports are simultaneously active for long periods of time. However, even a non-blocking switch may discard frames when a port becomes congested, depending on traffic patterns.

Packet Forwarding Performance

Typical switch hardware has dedicated support circuits that are designed to help improve the speed with which the switch can forward a frame and perform such essential functions as looking up frame addresses in the address filtering database. As support circuits and high-speed buffer memory are more expensive components, the total performance of a switch is determined by a trade-off between the cost of those high-performance components and the price most customers are willing to pay. Therefore, you will find that not all switches perform alike.

Some less expensive devices may have lower packet forwarding performance, smaller address filtering tables, and smaller buffer memories. Larger switches with more ports will typically have higher-performance components and a higher price tag. Switches capable of handling the maximum frame rate on all of their ports, also described as nonblocking switches, are capable of operating at *wire speed*. Fully nonblocking switches that can handle the maximum bit rate simultaneously on all ports are common these

3. For example, a 100 Mb/s Ethernet LAN can send a maximum of 148,809 frames per second when using the minimum frame size of 64 bytes.

days, but it's always a good idea to check the specifications for the switch you are considering.

The performance required and the cost of the switches you purchase can vary depending on their location in the network. The switches you use in the core of a network need to have enough resources to handle high traffic loads. That's because the core of the network is where the traffic from all stations on the network converges. Core switches need to have the resources to handle multiple conversations, high traffic loads, and long-duration traffic. On the other hand, the switches used at the edges of a network can be lower-performance, because they are only required to handle the traffic loads of the directly connected stations.

Switch Port Memory

All switches contain some high-speed buffer memory in which a frame is stored, however briefly, before being forwarded on to another port or ports of the switch. This mechanism is known as *store-and-forward switching*. All IEEE 802.1D–compliant switches operate in store-and-forward mode, in which the packet is fully received on a port and placed into high-speed port buffer memory (stored) before being forwarded. A larger amount of buffer memory allows a bridge to handle longer streams of back-to-back frames, giving the switch improved performance in the presence of bursts of traffic on the LAN. A common switch design is to provide a pool of high-speed buffer memory that can be dynamically allocated to individual switch ports as needed.

Switch CPU and RAM

Given that a switch is essentially a special-purpose computer, the central CPU and RAM in a switch are important for such functions as spanning tree operations, providing management information, managing multicast packet flows, and managing switch port and feature configuration.

As usual in the computer industry, the more CPU performance and RAM, the better, but you will pay more as well. Vendors frequently do not make it easy for customers to find switch CPU and RAM specifications. Typically, higher-cost switches will make this information available, but you won't be able to order a faster CPU or more RAM for a given switch. Instead, this is information useful for comparison between models from a vendor, or between vendors, to see which switches have the best specifications.

Switch Specifications

Switch performance includes a range of metrics, including the maximum bandwidth, or switching capacity of the packet switch electronics, inside the switch. You should also investigate the maximum number of MAC addresses that the address database can hold,

as well as the maximum rate in packets per second that the switch can forward on all ports.

Shown next is a set of switch specifications copied from a typical vendor's data sheet. The vendor's specifications are shown in italics. To keep things simple, in our example we show the specifications for a small, low-cost switch with five ports. This is intended to show you some typical switch values, and also to help you understand what the values mean and what happens when marketing and specifications meet on a single page.

Forwarding

Store-and-forward

Refers to standard 802.1D bridging, in which a packet is completely received on a port and into the port buffer ("stored") before being forwarded.[4]

128 KB on-chip packet buffering

The total amount of packet buffering available to all ports. The buffering is shared between the ports, on an on-demand basis. This is a typical level of buffering for a small, light-duty, five-port switch intended to support client connections in a home office.

Performance

Bandwidth: 10 Gb/s (non-blocking)

Because this switch can handle the full traffic load across all ports operating at the maximum traffic rate on each port, it is a non-blocking switch. The five ports can operate at up to 1 Gb/s each. In full-duplex mode, the maximum rate through the switch with all ports active can be 5 Gb/s in the outbound direction (also called "egress") and 5 Gb/s in the inbound direction (also called "ingress"). Vendors like to list a total of 10 Gb/s aggregate bandwidth in their specifications, although the 5 Gb/s of ingress data on five ports is being sent as 5 Gb/s of egress data. If you regarded the maximum aggregate data transfer through the switch as 5 Gb/s, you would be technically correct, but you would not succeed in marketing.[5]

4. Some switches designed for use in data centers and other specialized networks support a mode of operation called *cut-through switching*, in which the packet forwarding process begins before the entire packet is read into buffer memory. The goal is to reduce the time required to forward a packet through the switch. This method also forwards packets with errors, because it begins forwarding a packet before the error checking field is received.

5. If switch vendors marketed automobiles, then presumably they would market a car with a speedometer topping out at 120 mph as being a vehicle that provides an aggregate speed of 480 mph, as each of the four wheels can reach 120 mph at the same time. This is known as "marketing math" in the network marketplace.

Forwarding rate

10 Mb/s port: 14,800 packets/sec

100 Mb/s port: 148,800 packets/sec

1000 Mb/s port: 1,480,000 packets/sec

These specifications show that the ports can handle the full packet switching rate consisting of minimum-sized Ethernet frames (64 bytes), which is as fast as the packet rate can go at the smallest frame size. Larger frames will have a lower packet rate per second, so this is the peak performance specification for an Ethernet switch. This shows that the switch can support the maximum packet rate on all ports at all supported speeds.

Latency (using 1500-byte packets)

10 Mb/s: 30 microseconds (max)

100 Mb/s: 6 microseconds (max)

1000 Mb/s: 4 microseconds (max)

This is the amount of time it takes to move an Ethernet frame from the receiving port to the transmitting port, assuming that the transmitting port is available and not busy transmitting some other frame. It is a measure of the internal switching delay imposed by the switch electronics. This measurement is also shown as 30 μs, using the Greek "mu" character to indicate "micro." A microsecond is one millionth of a second, and 30 millionths of a second latency on 10 Mb/s ports is a reasonable value for a low-cost switch. When comparing switches, a lower value is better. More expensive switches typically provide lower latency.

MAC address database: 4,000

This switch can support up to 4,000 unique station addresses in its address database. This should be more than enough for a five-port switch intended for home office and small office use.

Mean time between failures (MTBF): >1 million hours (~114 years)

The MTBF is high because this switch is small, has no fan that can wear out, and has a low component count; there aren't many elements that can fail. This doesn't mean that the switch can't fail, but there are few failures in these electronics, resulting in a large mean time between failures for this switch design.[6]

6. Good luck returning the switch for a refund if it fails after, say, 70 years. May you live so long, but the vendor probably won't.

Standards compliance
IEEE 802.3i 10BASE-T Ethernet
IEEE 802.3u 100BASE-TX Fast Ethernet
IEEE 802.3ab 1000BASE-T Gigabit Ethernet
Honors IEEE 802.1p and DSCP priority tags
Jumbo frame: up to 9,720 bytes

Under the heading of "Standards compliance" the vendor has provided a laundry list of the standards for which this switch can claim compliance. The first three items mean that the switch ports support twisted-pair Ethernet standards for 10/100/1000 Mb/s speeds. These speeds are automatically selected in interaction with the client connection, using the Ethernet Auto-Negotiation protocol. Next, the vendor states that this switch will honor Class of Service priority tags on an Ethernet frame by discarding traffic with lower-priority tags first in the event of port congestion. The last item in this laundry list notes that the switch can handle nonstandard Ethernet frame sizes, often called "jumbo frames," which are sometimes configured on the Ethernet interfaces for a specific group of clients and their server(s) in an attempt to improve performance.[7]

This set of vendor specifications shows you what port speeds the switch supports and gives you an idea of how well the switch will perform in your system. When buying larger and higher-performance switches intended for use in the core of a network, there are other switch specifications that you should consider. These include support for extra features like multicast management protocols, command-line access to allow you to configure the switch, and the Simple Network Management Protocol to enable you to monitor the switch's operation and performance.

When using switches, you need to keep your network traffic requirements in mind. For example, if your network includes high-performance clients that place demands on a single server or set of servers, then whatever switch you use must have enough internal switching performance, high enough port speeds and uplink speeds, and sufficient port buffers to handle the task. In general, the higher-cost switches with high-performance switching fabrics also have good buffering levels, but you need to read the specifications carefully and compare different vendors to ensure that you are getting the best switch for the job.

7. Jumbo frames can be made to work locally, for a specific set of machines that you manage and configure. However, the Internet consists of billions of Ethernet ports, operating with the standard maximum frame size of 1,500 bytes. If you want things to work well over the Internet, stick with standard frame sizes.

Basic Switch Features

Now that we've seen how switches function, we will describe some of the features you may find supported on switches. The size of your network and its expected growth will affect the way you use Ethernet switches and the types of switch features that you need. A network in a home or single office space can get by with one or a few small and low-cost switches that provide basic Ethernet service at high enough speeds to meet your needs and few extra features. Such networks are not expected to be complex enough to present major challenges in terms of network stability, nor are they expected to grow much larger.

On the other hand, a medium-sized network supporting multiple offices may need more powerful switches with some management features and configuration capabilities. If the offices require high-performance networking for access to file servers, then the network design may require switches with fast uplink ports. Large campus networks with hundreds or even thousands of network connections will typically have a hierarchical network design based on switches with high-speed uplink ports, with more sophisticated switch features to support network management and help maintain network stability.

Switch Management

Depending on their cost, switches may be provided with a management interface and management software that collects and displays statistics on switch operation, network activity and port traffic, and error. Many medium- and higher-cost switches include some level of management capability, and vendors typically provide management application software that is Web-based and that may also allow you to log into the switch via a console port on the switch or over the network.

The management software allows you to configure port speeds and features on the switch; it also provides monitoring information on switch operations and performance. Switches that support the Spanning Tree Protocol typically also support a management interface that allows you to configure spanning tree operations on each switch port. Other configurable options may include port speed, Ethernet auto-negotiation features, and any advanced switch features that may be supported.

Simple Network Management Protocol

Many switch management systems use the Simple Network Management Protocol (SNMP) to provide a vendor-neutral way to extract operational information from a switch and deliver that data to you. That information typically includes the traffic rates being seen on switch ports, error counters that can identify devices that are having problems, and much more. Network management packages based on SNMP protocols can retrieve information from a wider range of network equipment than just switches.

There are multiple software packages available in the marketplace that can retrieve SNMP-based management information from a switch and display it to the network manager. There are also a number of open source packages that provide access to SNMP information and display that information in graphs and textual displays. See Chapter 21 for links to further information.

Packet Mirror Ports

Another useful feature for monitoring and troubleshooting switches is called a *packet mirror port*. This feature allows you to copy, or "mirror," the traffic from one or more ports on the switch to the mirror port. A laptop running a network analyzer application can be connected to the mirror port to provide network traffic analysis.

A mirror port can be a very useful feature for tracking down network problems on devices connected to a given switch. Vendors have adopted a wide range of approaches to mirror ports, with different capabilities and limitations depending on their particular implementation. Some vendors even make it possible for mirrored traffic to be sent to a remote receiver over the network, which enables remote troubleshooting. Packet mirroring ports are not a standardized feature of switches, so vendors may or may not include support for this feature.

Switch Traffic Filters

Switch traffic filters make it possible for a network manager to specify Ethernet frame filtering based on a number of parameters. The range of filters supported by switches varies widely among vendors. Lower-cost devices with no management interface won't have any filtering capability, while higher-cost and higher-performance devices may offer a complete set of filters that the network manager can set.

By using these filters, a network manager can configure switches to control network traffic based on such things as the addresses of Ethernet frames and the type of high-level protocol being carried in the frames. Filters may result in reduced performance, so you should check the switch documentation to determine the impact.

Filters work by comparing filter patterns, expressed as numeric values or protocol port names (e.g., *http*, *ssh*), against the bit patterns seen in Ethernet frames. When the pattern matches, then the filter takes some action, typically dropping the frame and thereby blocking the traffic. Be aware that by using filters, you may cause as many problems as you are trying to resolve.

Filters that are designed to match patterns in the data field of the frame can cause issues when those patterns also occur in frames that you did not want to filter. A filter set up to match on one set of hex digits at a given location in the data field of a frame may work fine for the network protocol you are trying to control, but could also block a network protocol you didn't even know existed.

This kind of filter is typically deployed to control the flow of some network protocol by identifying a part of the protocol in the data field of the Ethernet frame. Unfortunately, it's hard for a network manager to anticipate the range of data that the network may carry, and depending on how it was constructed, the filter may match frames that were not intended to be filtered. Debugging a failure caused by a wayward filter can be difficult, because it's usually not very obvious why an otherwise normally functioning Ethernet stops working for a specific application or for a certain set of stations.

Switch filters are often used in an attempt at greater control by preventing network interaction at the high-level network protocol layer of operations. If that's why you're implementing switch filters, then you should consider using Layer 3 routers that operate at the network layer and automatically provide this level of isolation without the need for manually configured filters.

Layer 3 routers also provide filtering capabilities that can be easier to deploy because they are designed to work on high-level protocol fields and addresses. This makes it possible to easily write a filter that protects your network equipment from attack, for example, by limiting access to the TCP/IP management addresses of the equipment. We'll look further at the use of routers in the next chapter.

Managing switch filters

It can be a complex undertaking to set up filters correctly, as well as to maintain them once they're in place. As your network grows, you will need to keep track of which switches have filters in them and make sure that you can remember how the filters you have configured affect the operation of the network system, as it can often be difficult to predict the effects of a filter.

Documentation of the filters you have deployed and the way they are being used can help reduce troubleshooting time. However, no matter how well documented they are, these kinds of filters can cause outages. Therefore, you should regard the use of filters as something to be done only when necessary, and as carefully as possible.

Virtual LANs

A widely used feature typically included in higher-cost switches is the ability to group ports in a switch into virtual local area networks, or VLANs. At its simplest, a VLAN is a group of switch ports that function as though they were an independent switch. This is done by manipulating the frame forwarding software in the switch.

If the vendor supports VLANs on a switch, it will provide a management interface to allow the network manager to configure which ports belong to which VLANs.

As shown in Figure 18-6, you could configure an eight-port switch so that ports 1 through 4 are in one VLAN (call it VLAN 100) and ports 5 through 8 in another (call it VLAN 200). Packets can be sent from station 10 to station 20, but not from station 10

to stations 30 and 40. As these VLANs act as separate networks, a broadcast or multicast sent on VLAN 100 will not be transmitted on any ports belonging to VLAN 200—the VLANs behave as though you had split the eight-port switch into two independent four-port switches.

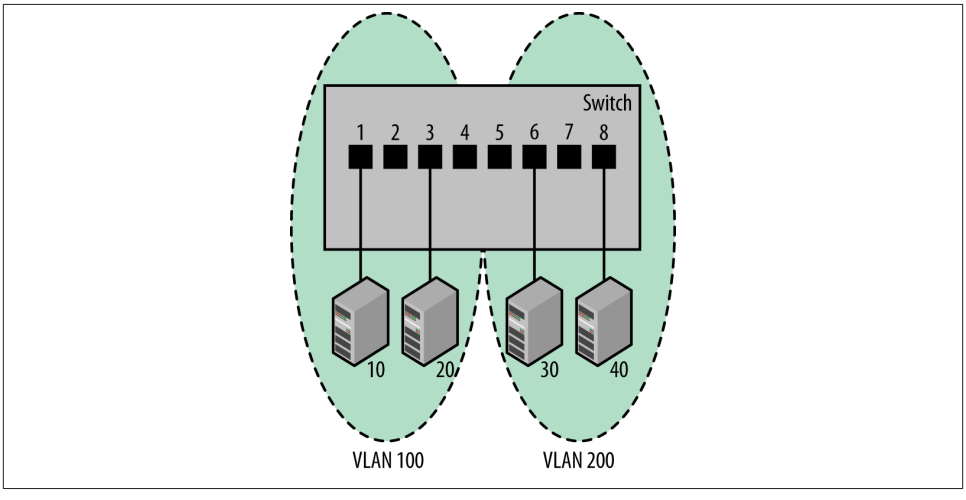


Figure 18-6. VLANs and switch ports

Vendors have provided other VLAN capabilities. For example, VLAN membership can be based on the contents of frames instead of just specifying which ports on the switch are members of a given VLAN. In this mode of operation, frames are passed through a set of filters as they are received on a switch port. The filters are set up to match some criterion, such as the source address in the frame or the contents of the type field, which specifies the high-level protocol carried in the data field of the frame. VLANs are defined that correspond to these filters; depending on which set of criteria the frames match, the frames are automatically placed into the corresponding VLAN.

802.1Q VLAN standard

The IEEE 802.1Q VLAN tagging standard was first published in 1998. This standard provides a vendor-independent way of implementing VLANs. As illustrated in Figure 18-7, the VLAN tagging scheme used in 802.1Q adds four bytes of information to the Ethernet frame, following the destination address and preceding the type/length field. This increases the maximum frame size in Ethernet to 1,522 bytes.

The 802.1Q standard also provides for priority handling of Ethernet frames using Class of Service (CoS) bits defined in the 802.1p standard. The 802.1Q standard provides space in the VLAN tag that allows you to use 802.1p CoS bits to indicate traffic priorities.

There are three bits reserved for CoS values, making it possible to provide eight values (0–7) to identify frames with varying service levels.

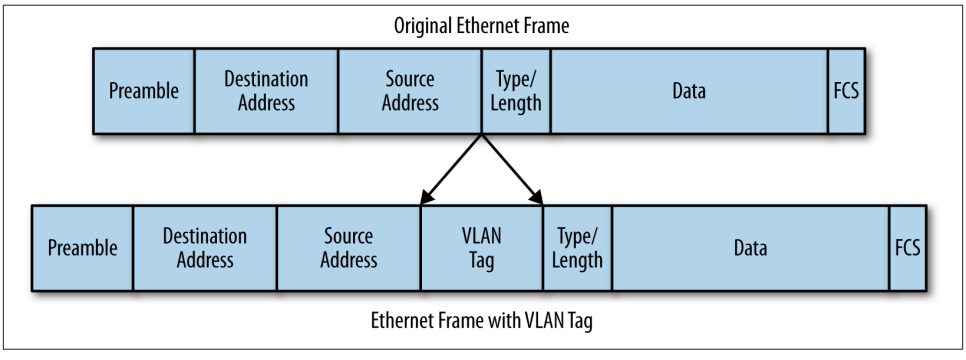


Figure 18-7. Ethernet frame with a VLAN tag

Linking VLANs

VLANs are widely used in network designs to provide multiple independent networks. This helps control the flow of traffic by isolating traffic sent among specific groups of stations. Each VLAN functions as a separate switch, and there is no way to link separate VLANs other than connecting them together with a Layer 3 router.

Building larger networks by linking VLANs with Layer 3 routers also avoids the propagation of broadcasts and multicasts seen on large Layer 2 networks, by shifting the packet forwarding operations between the VLANs to a Layer 3 protocol.

Floods of broadcasts and multicasts can occur due to network path loops, broken software, and address scanning attacks, among other reasons. All devices connected to a given VLAN will be subjected to the traffic flood, which can also be described as all devices being in the same "failure domain." Creating separate VLANs and linking them with Layer 3 routers also creates separate failure domains. By connecting the population of devices on separate VLANs you can thereby limit the number of devices in a given failure domain, and improve reliability, availability, and network stability.

802.1Q Multiple Spanning Tree Protocol

The Multiple Spanning Tree Protocol (MSTP) was developed in the 802.1s supplement in 2003 and was included in the 2005 edition of the 802.1Q standard. It is based on RSTP and is defined as an optional extension to RSTP to add the ability for switches supporting VLANs to use multiple spanning trees.

This makes it possible for traffic belonging to different VLANs to flow over different paths within the same network of switches. Thus, the MSTP standard is "VLAN aware" and is well suited for operation in networks with many VLANs and multiple uplink

paths. The operation of the MSTP standard was also designed to minimize the number of BPDUs required to build spanning trees for multiple VLANs, and to therefore be a more efficient system.

Note that the classic STP and the more recent RSTP are sufficient for many networks. Even when there are VLANs in a network, these protcols will still be able to block loop paths. MSTP was developed to provide VLAN awareness for more complex network designs, and to provide a more efficient model of operation based on multiple spanning tree (MST) "regions" that can each run a number of MST instances (up to 64). The use of multiple regions requires that the network administrator configure MST bridges to be members of specific regions, potentially making MST more complex to set up and operate.

While the MST standard provides advantages in terms of structuring a large system into regions and reducing the processing required to maintain the spanning tree, it can also require more up-front effort to understand the configuration requirements and to implement them in your switches. Vendors are adopting MSTP as the default spanning tree system on some switches—typically high-performance systems intended for use in data centers and capable of supporting large numbers of VLANs. However, RSTP, and even the classic STP, are still widely used versions of spanning tree. Their "plug and play" operation, with its ability to create a spanning tree without any configuration effort, is effective for many network designs.

Quality of Service (QoS)

Managing the priority of traffic flow to favor certain categories of traffic over other categories when congestion occurs is another capability of switches. The 32-bit field added by the IEEE 802.1Q standard provides support for traffic prioritization fields to provide eight different Class of Service values as well as the VLAN tag.

The 802.1p standard provides traffic prioritization levels that are carried in the 802.1Q CoS tag and used to tag a frame with a priority value, so that certain traffic may be favored for transmission when network congestion occurs on a switch port. When CoS has been configured on a switch port, then the Ethernet frames that are not tagged with a high priority are the first to be dropped should congestion occur on the port.

If your switch supports these features, you will need to consult the vendor documentation for instructions on how to configure them. While the IEEE standards describe the mechanisms that make these features possible, the standards do not specify how they should be implemented or configured. That's left up to each vendor to decide, which means that the vendor documentation is the place to find the details on how to use these features in a given switch.

CHAPTER 19

Network Design with Ethernet Switches

In this chapter, we will show some of the basic ways in which switches can be used to build an Ethernet system. Network design is a large topic, and there are many ways to use switches to expand and improve networks. We will focus on just a few basic designs here, with the goal of providing a brief introduction to network design with Ethernet switches.

Advantages of Switches in Network Designs

Switches provide multiple advantages in network designs. To begin with, all switches provide the basic traffic filtering functions described in the previous chapter, which improves network bandwidth. Another important advantage of modern switches is that the internal switching electronics allow different traffic flows to occur simultaneously between multiple ports. Supporting multiple simultaneous flows of traffic, or "conversations," between the ports is a major advantage of switches in network designs.

Improved Network Performance

An important way in which a switch can improve the operation of a network system is by controlling the flow of traffic. The ability to intelligently forward traffic on only those ports needed to get the packet to its destination makes the switch a useful tool for the Ethernet designer faced with continually growing device populations and increasing traffic loads.

The traffic control provided by the internal address database can be exploited to help isolate traffic. By locating client and server connections on switches to help minimize network traffic, you can keep the traffic between a set of clients and their file server localized to the ports on a single switch. This keeps their traffic from having to traverse the larger network system.

Figure 19-1 shows a set of clients and their file server connected to a single switch, Switch 2, which isolates their traffic from the rest of the network connections in the building. In this design, all of the local traffic between clients 40, 50, and 60 and their file server stays on Switch 2 and does not travel through any other switches in the building.

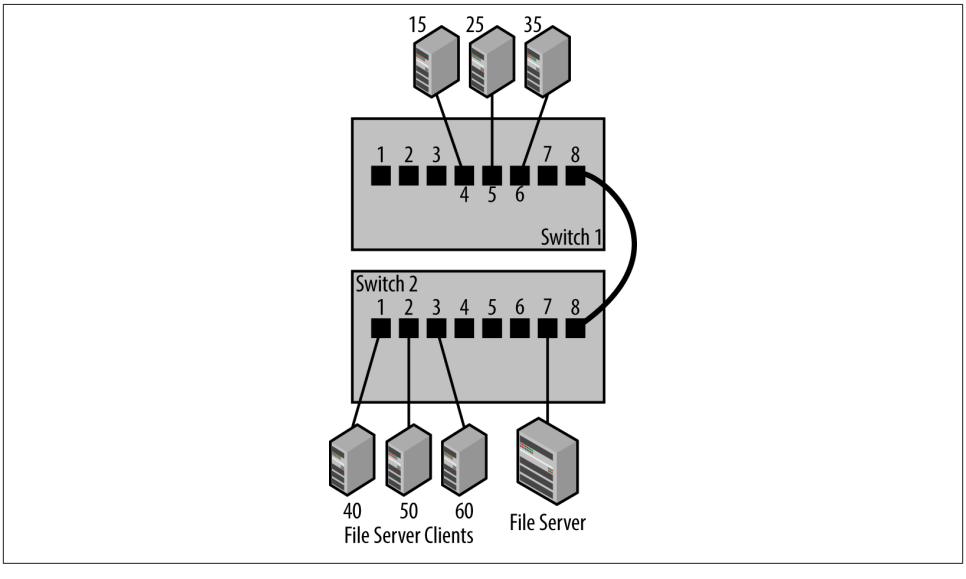


Figure 19-1. Isolating client/server traffic

When installing a switch, you can improve your network's operation by being aware of the traffic flows and designing the network so that the traffic between a cluster of clients and their server(s) stays local. You may not be able to do this for all clients in a building, but any clients and servers that you can keep on a single switch (or small group of switches), will help minimize the amount of traffic that has to cross all switches in your network.

This example reveals another important issue, which is that the links used to connect switches together should be high-performance links. Links between switches are called *uplinks*, because network tree diagrams are typically drawn with the switches in a top-to-bottom hierarchy. The topmost switch is the *core* switch, which functions as the core of the network system by linking all the other switches.

Linking the *edge* switches directly to the core in this fashion minimizes the number of switches that the network traffic must cross (also known as *switch hops*) to get from one computer to another in your network. Uplinks connect one switch to the next, leading up to a higher level of the network (toward the core). Traffic travels in both directions over the uplinks.

Switch Hierarchy and Uplink Speeds

Another advantage of switches is that they can link multiple network connections that run at different speeds. Any given connection to a switch port runs at a single speed. However, multiple computers can be connected to the same switch, with the connections operating at different speeds. Depending on its cost and feature set, you may find that your switch has a couple of ports described as *uplink ports*. These ports typically support higher speeds than the rest of the ports on the switch and are intended for making a connection up to the core switches, hence the "uplink" name.[1]

Switch ports can run at different speeds because a switch is equipped with multiple Ethernet interfaces, each capable of operating at any of the speeds supported on the interface. A switch can read in an Ethernet frame on a port operating at 1 Gb/s, store the frame in port buffer memory, and then send the frame out on a port operating at 10 Gb/s.

Filling the port buffer and causing congestion and dropped frames is more likely to occur when receiving on a 10 Gb/s port and sending on a 1 Gb/s port, due to the large difference in speeds and the longer time it takes to send a frame out the 1 Gb/s port.

In Figure 19-2, three edge switches are shown, each with one of its uplink ports connected to a fourth switch located at the core of the network. While the uplink ports operate at 10 Gb/s, most of the station connections are at 1 Gb/s, except for the file server, which is connected to a 10 Gb/s port.

This connection shows that it's possible to connect the server to one of the uplink ports, because there's nothing that prohibits an uplink port from operating as a station port. Uplink ports typically operate at higher speeds, and they typically have a larger buffer memory to handle traffic arriving at higher speeds on the uplink port (10 Gb/s) that is destined for a slower station port (1 Gb/s). For that reason, while you usually want to save these ports for uplink purposes, they can also be used to connect to a heavily used server machine.

1. If you want to use the latest network jargon, you could say that the uplink ports are used to create "northbound" connections to the core of your network.

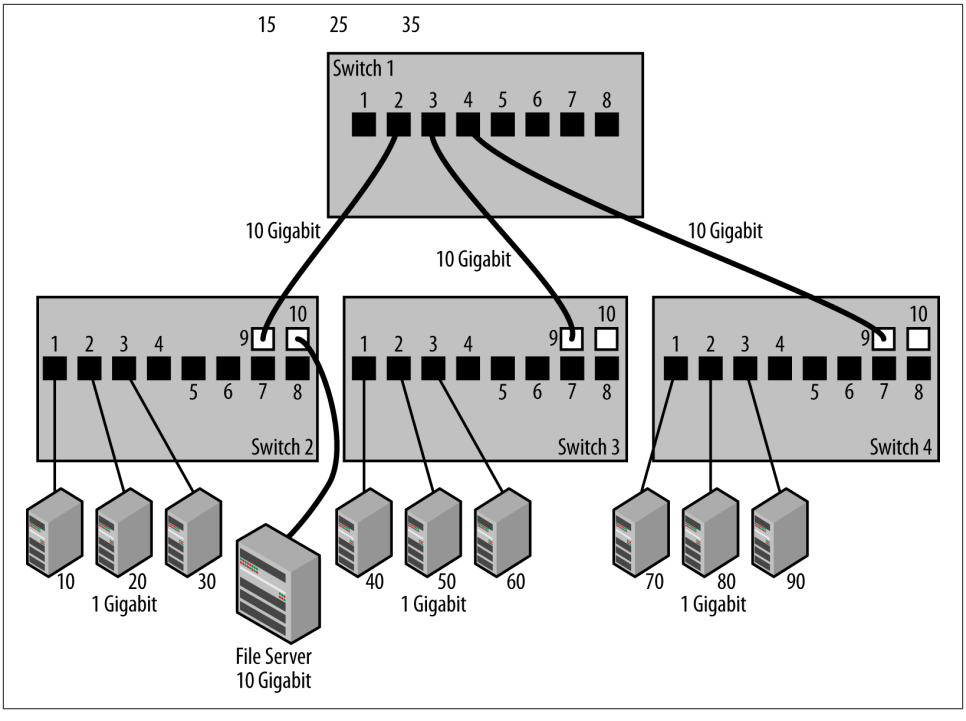


Figure 19-2. Switch hierarchy and uplink speeds

Uplink Speeds and Traffic Congestion

The main reason you want uplinks to run faster is that there may be traffic from multiple stations connected at 1 Gb/s attempting to get to a single server on a separate switch. If the uplink connections were also running at 1 Gb/s, then there could be congestion and dropped frames, which can cause reduced performance for the computers sending data.

The port buffers on a switch are deliberately designed to hold just a few packets for a short period, to allow for a small amount of congestion. If the buffers were too large, there could be increased packet delay and variation in packet delivery times, causing problems for certain kinds of applications.

If there is a lot of traffic continually being sent from clients into one or more congested ports, the switch will run out of space on the port buffer(s) and will simply drop the incoming packets until space becomes available.

Local area networks are not designed to provide guaranteed packet delivery. If the system becomes congested, packets are dropped. The TCP/IP network protocol, in turn, is designed to respond to dropped frames by automatically throttling the traffic rate. In other words, dropped frames are normal and in fact are required to allow TCP/IP to detect and respond to network congestion.

To see how this works, let's take the example shown in Figure 19-2. Suppose that the three stations 70, 80, and 90 on Switch 4 all need to retrieve files from the file server on Switch 2. File service traffic tends to be high-bandwidth; you could easily end up with three 1 Gb/s streams from the file server to the stations on Switch 4. If the uplink connections are operating at 1 Gb/s, then the switch ports in the path between the stations on Switch 4 and the file server on Switch 2 will become congested and will drop the frames that cannot fit into the already-full port buffers.

If, on the other hand, you link the switches together with 10 Gb/s uplinks, then you have a 10 Gb/s path from the file server on Switch 2 into Switch 4, and all three stations on Switch 4 will be able to interact with the file server at their maximum network rate of 1 Gb/s, without causing major congestion on the uplink paths. Packets received at 1 Gb/s from the stations will be sent 10 times as fast, at 10 Gb/s, over the uplinks; this rapidly drains the uplink port buffers and ensures that there is buffer space available for more traffic from stations. Another possible design is to connect the server directly to the core switch on a 10 Gb/s port.

Multiple Conversations

The connection method shown here for the uplinks illustrates a couple of major advantages of switches: improved packet switching performance and support for multiple simultaneous traffic flows between stations and the file server. Every port on the switch in our example is an independent network connection, and each station gets its own 1 Gb/s dedicated full-duplex Ethernet path directly into the switch. Multiple station conversations can be going on simultaneously in both directions (data from the computer and replies to the computer), providing high performance and minimizing network delay.

Referring back to Figure 19-2, this means that while station 70 and the file server are communicating, station 80 and station 10 also can be communicating at the same time. In this configuration, the total network bandwidth available to stations becomes a function of the ports to which each station is connected, and of the total packet switching capacity of your switch. Modern switches are equipped with switching fabrics inside the switch that can provide many gigabits per second of switching capacity, and high-end switches will provide up to multiple terabits of packet switching capacity.

The speed of the switching fabric is only one important consideration when it comes to moving frames through a switch. As we've seen, high traffic rates coming into the switch from multiple ports, all destined for a single server port on the switch, will always be an issue, because the server port can only deliver packets at its maximum bit rate, no matter how many packets are sent to it at any given moment.

Switch Traffic Bottlenecks

When multiple flows occur, it's possible to overrun the capability of the output port, no matter how much internal packet switching capacity the switch may have. The switch will start dropping frames when it runs out of buffer space to store them temporarily. A dropped frame causes the network protocol software running on the computer to detect the loss and retransmit the data. Too many data retransmissions caused by excessive congestion on an output port can lead to a slow response for the application that is trying to talk to the server.

Traffic bottlenecks such as these are an issue in all network designs. When linking switches together, you may encounter situations where a bottleneck occurs when the traffic from multiple switches must all travel over single backbone links connecting two core switches. If there are multiple parallel connections linking the core switches, the spanning tree algorithm will ensure that only one path is active, to prevent loops in the network. Therefore, the ports of the switches that feed the single inter-core-switch link could be facing the same situation as the oversubscribed server port just described, causing the core switch ports to drop frames. In sufficiently large and busy network systems, a single interswitch link may not provide enough bandwidth, leading to congestion.

There are several approaches that can be taken to avoid these problems in network systems. For example, the IEEE 802.1AX link aggregation standard allows multiple parallel Ethernet links to be grouped together and used as a large "virtual" channel between backbone switches.

Link aggregation was first defined in the IEEE 802.3ad standard, and then later moved to become 802.1AX. You will find both standards referred to in vendor documentation.

Using link aggregation, multiple Gigabit Ethernet links can be aggregated into channels operating at 2, 4, and 8 gigabits per second. The same is true for 10 Gigabit links, providing a channel operating at up to 80 gigabits per second. This approach can also be used between a switch and Ethernet interfaces in high-performance servers to increase network bandwidth.

Another approach is to use Layer 3 routers instead of Layer 2 switches, because routers don't use the spanning tree algorithm. Instead, routers provide more sophisticated traffic routing mechanisms that make it possible for network designers to provide multiple parallel connections for backbone links that are simultaneously active.

Hierarchical Network Design

Network design refers to the way that switches are interconnected to produce a larger network system. Any network with more than just a few switches, and especially any network that is expected to grow, will benefit from a hierarchical network design that results in a system that is higher-performance, more reliable, and easier to troubleshoot. Implementing a network design, and thus providing a plan for optimizing network operation and growth, pays major dividends in terms of network performance and reliability.

Networks that grow without a plan often result in systems with switches connected together in such a way that there are more switches in the traffic paths than necessary. This, in turn, leads to more complex "mesh" designs that are harder to understand and troubleshoot.[2]

Systems that "just grew" may also develop traffic bottlenecks whose presence and location are a mystery to the network administrator.

It's a fact of life that networks often grow; without an adequate design in place they will grow randomly, becoming ever more complex and difficult to understand. A simple hierarchical network design based on two or three "layers" minimizes the number of switches needed, improving traffic flow and resulting in improved network performance and reliability. Other important advantages are that the network will be more stable and understandable as the system grows over time.

The most widely deployed design for networks that support standard offices and cube space in a campus of buildings is based on a hierarchical system of three layers: the *core*, *distribution*, and *access layers* (as shown in Figure 19-3). The core layer contains the high-performance switches that connect all buildings on a campus together. Each building has a distribution point, which contains the medium-performance switches connecting the building to the core and also connecting to the access switches inside the building. Finally, there is an access layer of switches, connecting to all devices in the building and connected in turn to the distribution switches. If there is only a single building, then the distribution and access layers are all that is needed, with the distribution layer functioning as the building core.

2. Another name for the result of network growth with no plan is "fur ball" design. (Or perhaps it should be called "hair ball.")

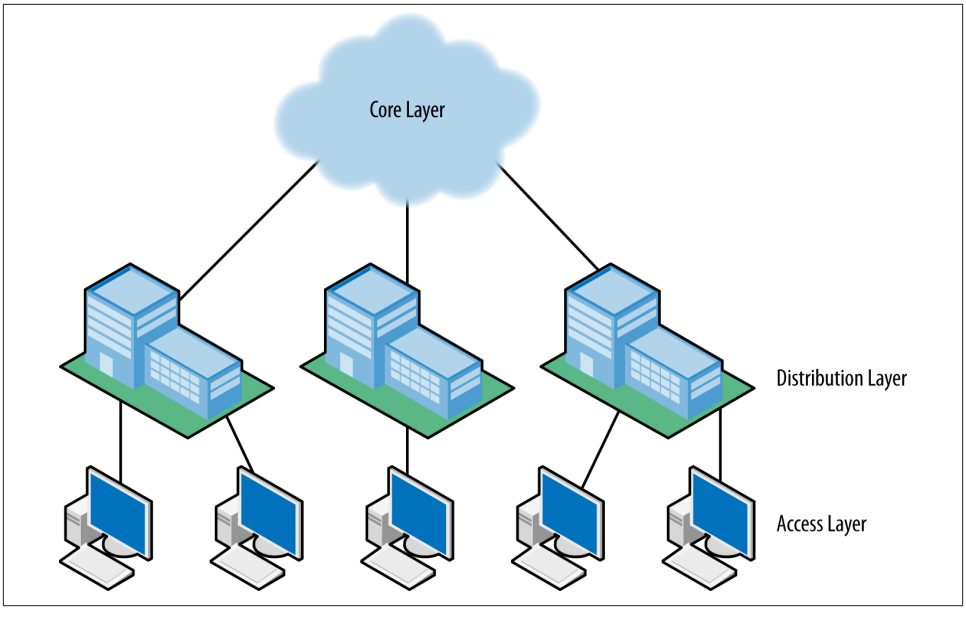


Figure 19-3. Hierarchical network architecture

Inside each building, the access switches are each connected directly to the distribution layer, and *not* to each other. It is essential that the uplinks of the access switches are connected only to the distribution-layer switches, to avoid creating a set of horizontal paths between the access switches with a much more complex mesh structure. A major benefit of this design, illustrated in Figure 19-4, is that it reduces the number of switches in the network path between communicating Ethernet devices. That, in turn, decreases the impact of switch latency, and also reduces the number of bottlenecks affecting network traffic.

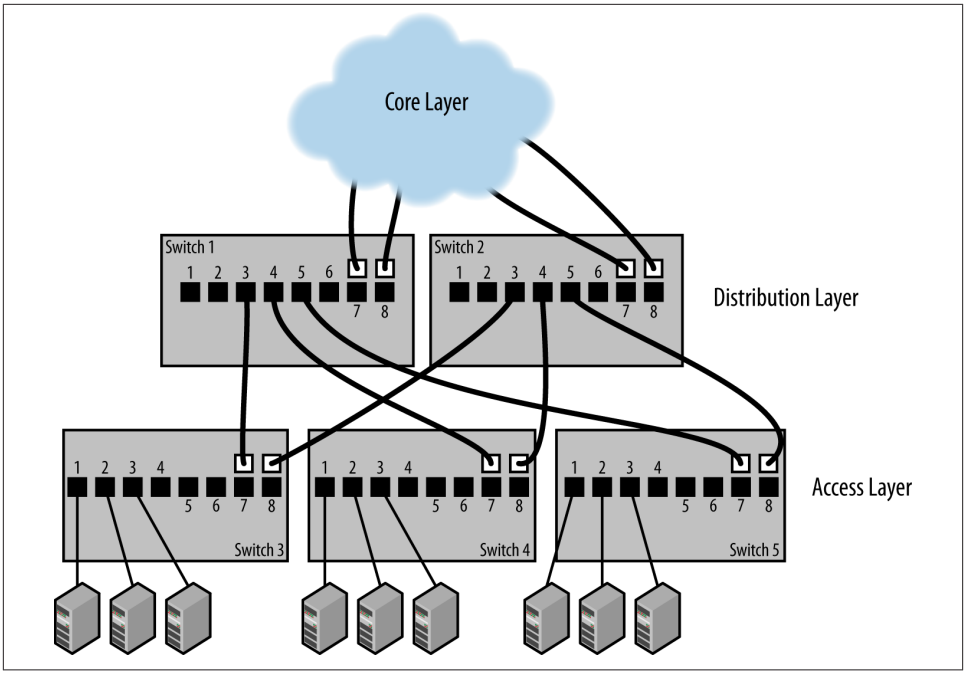


Figure 19-4. Distribution network in a building

This design also minimizes the number of potential loop paths, which helps the Spanning Tree Protocol converge on a set of paths more rapidly (see Chapter 18 for more on loops and the STP). This can become especially important after a power failure in a complex network, when all switches come back up at the same time and the Spanning Tree Protocol must work to stabilize the entire set of network paths simultaneously. A hierarchical network design also makes it easier to provide high-bandwidth uplinks, which helps avoid major bottlenecks and keeps spanning tree operations working well under all network load conditions.

Establishing and maintaining a network design requires documentation and attention to detail. Everyone involved in maintaining the network needs to understand the design in use and the benefits of maintaining a good network structure. You will find pointers to further reading on network design in Appendix A.

Seven-hop maximum

As we've just seen, there are multiple reasons for minimizing the number of switches in the network path between devices. The 802.1D bridging standard provided yet an-

other reason when it recommended a maximum network diameter of seven hops, meaning seven switches in the path between any two stations.[3]

The recommended limit on the number of switches originated from a concern about round-trip packet delays with 7 switches in a given path, providing 14 switch hops in a total round trip. A time-sensitive application sending a frame from one end of the network to the other and receiving a reply would encounter 14 switch hops, with the potential for an impact on the performance of the application because of the time required to transition 14 switches.

Subsequent versions of the standard removed the seven-hop recommendation from the standard. However, in large network designs, an important goal of your network design should be to keep the total number of Layer 2 switch hops to a minimum.

Network Resiliency with Switches

Network systems support access both to the Internet and to all manner of local computing resources, making them essential to everyone's productivity. If the network fails, it will have a major impact on everyone's ability to get their work done. Fortunately, network equipment tends to be highly reliable, and equipment failures are rare. Having said that, network equipment is just another computer in a box. There are no perfect machines; at some point you *will* have a switch failure. The power supply may quit working, causing the entire switch to fail, or one or more ports may fail. If an uplink port fails, it could isolate an entire downstream switch, cutting off network access for every station connected to that switch.

One way to avoid network outages due to a switch failure is to build resilient networks based on the use of multiple switches. You could purchase two core switches, call them Switch 1 and Switch 2, and connect them together over parallel paths, so that there are two links between them in case one of the links fails. Next, you could link each of the access switches that connect to stations to both core switches. In other words, on each access-layer switch, one of the two uplink ports would be connected to core Switch 1 and the other is connected to core Switch 2.

Figure 19-5 shows the two core switches, Switch 1 and Switch 2, connected together over two parallel paths to provide a resilient connection in case one of the links fails for any reason. The aggregation switches are each connected to both core switches, providing two paths to the network core in case any single path fails.

3. The seven-hop maximum limit recommendation was in all versions of the 802.1D standard up to 1998.

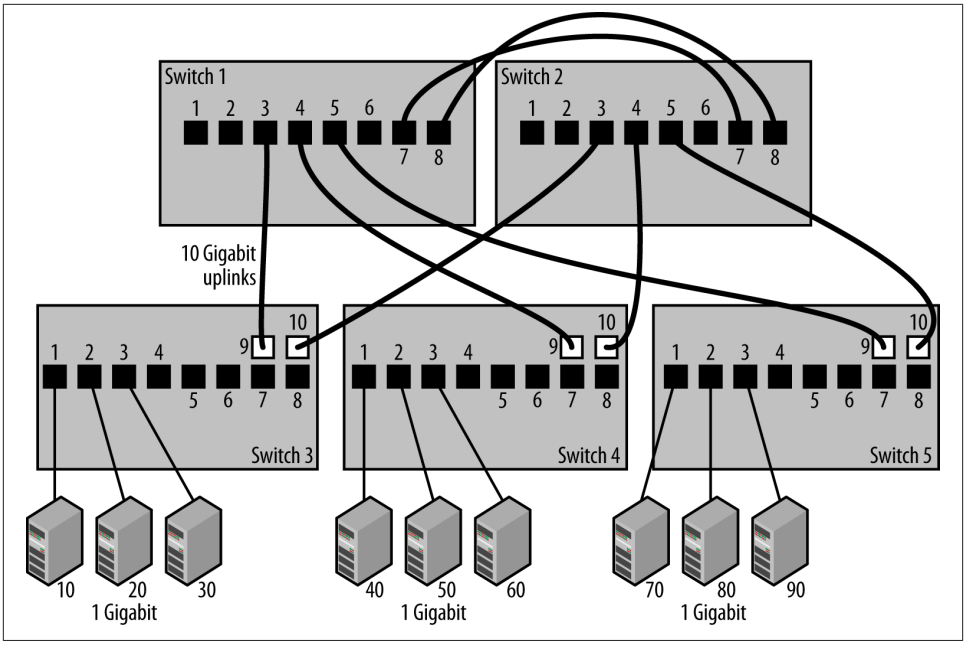


Figure 19-5. Network resiliency with switches

Spanning Tree and Network Resiliency

At this point, you should be asking, "But what about spanning tree? Won't it shut down those parallel paths between resilient switches?" The answer is yes, spanning tree will block one of the two paths to ensure that there are no loop paths in the network system. The path will stay blocked until one of the active links fails, in which case RTSP, responding quickly to a detected change in the network, will immediately bring the backup path into operation.

Figure 19-6 shows the resilient design after spanning tree has suppressed the loop paths by blocking the forwarding of packets on certain uplink ports. The ports that are blocked are shown with a B. If you know the MAC addresses and bridge IDs for each switch, then you can calculate, based on the operation of the Spanning Tree Protocol, exactly which ports will be blocked to prevent loop paths.

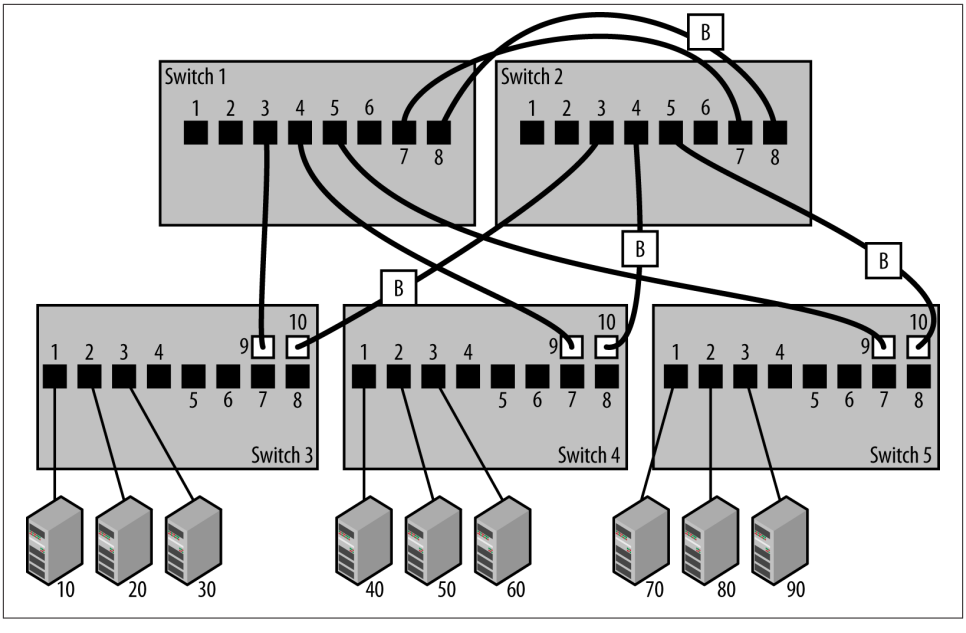


Figure 19-6. Spanning tree suppresses loop paths

But you don't need to know how things work at that level of detail. Instead, spanning tree will function automatically and block one of the two paths creating a loop. The Ethernet link stays up and connected on the blocked port, but no traffic is forwarded over it. If the remaining active path should ever fail, then spanning tree will automatically reenable the blocked port to bring the link back into operation.

Cost and complexity of resiliency

The combination of dual-core switches, dual uplinks, and spanning tree can provide a resilient design that can survive the failure of one of the core switches and/or one of the dual uplinks from the aggregation-layer switches. However, this is a higher-cost and higher-complexity design that requires both a larger budget and an understanding of how to connect switches for resilience. This design also has the disadvantage of providing only resiliency. All of the access switch links to Switch 2 are in the blocking state, and Switch 2 does not carry any traffic unless there is a failure of Switch 1.

If your network uptime needs are not so stringent as to require the kind of high uptime and automatic failure recovery provided by this design, then you could get by with keeping a spare switch on hand and using it to replace failures when they occur. It's all a matter of your tolerance for network downtime and how much you are willing to invest in avoiding outages with automatic recovery systems.

Network design issues such as these require a good deal of knowledge about the mechanisms used to direct the flow of traffic through various devices such as switches and routers. This is also an area that is undergoing rapid evolution, and new mechanisms for moving traffic around large networks are continually being invented and tried out in the marketplace.

Routers

A router is a device that operates at the network layer (Layer 3) of the OSI reference model. It helps to understand that the OSI layers are not derived from physics or the natural laws of the universe. Quite the opposite. Instead, the OSI layers are arbitrary definitions used to group the various details involved in computer communications into associated tasks called "layers." This was done as a way to help clarify the tasks and to structure the development of the standards needed to achieve the communication tasks.

Like many other human endeavors, the evolution of computer communications technology has not followed a completely logical development path. For example, local area networks were defined as operating at Layer 2 because that's what the people developing the standards wanted: a local network that carried data between computers located at a given site. Layer 2 standards describe local area networks operating at the data link layer, and were not intended to deal with the issues of interconnecting large numbers of networks.

More sophisticated protocol operations based on structured addresses and capable of dealing with large numbers of networks were defined at Layer 3, the network layer. Switches operate at Layer 2 using only the information found in Ethernet frames, and routers operate at Layer 3 using high-level network protocol packets carried in the data field of Ethernet frames, such as those packets defined in the TCP/IP protocol suite. Both Layer 2 and Layer 3 switches use Ethernet frames, but the addressing information used by the respective switches to make packet forwarding decisions is very different.

Operation and Use of Routers

Routers are frequently used in large campus and enterprise networks, as well as the worldwide Internet. At the network layer of operation, you can find a wider range of mechanisms for interconnecting large network systems. While routers are more complex to configure than switches, the advantages they can provide offset the added complexity of their operation for many network managers.

In operation, a router receives and unpacks the Ethernet frames, and then applies rules to deal appropriately with the high-level protocol data that is carried in the data field of each Ethernet frame. When a router hears an Ethernet broadcast, it does the same thing all other stations must do: it reads in the frame and tries to figure out what to do with it.

Routers are moving packets around based on higher-level protocol addresses so they do not forward Ethernet broadcast or multicast packets. Broadcast or multicast packets sent from a client station attempting to discover a service on a server connected to the same local network, for example, are not forwarded to other networks by the router, because the router is not designed to create a larger local area network.

Dropping broadcasts and multicasts at the router interface has the effect of creating separate broadcast domains, protecting a large network system from the high multicast and broadcast traffic rates that might otherwise occur. This is a major advantage, both for the reduced traffic levels and for the reduction in computer performance issues that can be caused by floods of broadcast and multicast packets.

Dividing networks into multiple smaller Layer 2 networks by linking them with Layer 3 routers also improves reliability by limiting the size of the *failure domain*. In the event of such failures as packet floods caused by loop paths, a failing station that is sending continuous broadcast or multicast traffic, or hardware or software failure resulting in a failures of spanning tree, the size of the failure domain is limited by using Layer 3 routers to link networks.

However, creating smaller Layer 2 networks and linking them with Layer 3 routers also limits the number of stations that can interact when using discovery services based on Layer 2 multicast and broadcast. That, in turn, may cause challenges for network designers who are attempting to grow their networks and limit the size of their failure domains while also keeping their users happy.

Users are happiest when everything "just works," and are often insistent on large Layer 2 networks to keep automatic service discovery working for the largest number of computers. However, when a large Layer 2 network fails, users may suddenly discover that network reliability is more important to them than widespread Layer 2 service discovery. Appendix A contains pointers to resources for more information on network design issues.

Routers or Bridges?

Although both Layer 2 switches (bridges) and Layer 3 switches (routers) can be used to extend Ethernets and build larger network systems, bridges and routers operate in very different ways. It's up to you to decide which device is best suited to your needs, and which set of capabilities is most important for your network design. Both bridges and routers have advantages and disadvantages.

Advantages of using bridges include the following:

- Bridges may provide larger amounts of switching bandwidth and more ports for lower cost than a router.

- Bridges may operate faster than a router, because they provide fewer functions.
- Bridges are typically simpler to install and operate.
- Bridges are transparent to the operation of an Ethernet.
- Bridges provide automatic network traffic isolation (except for broadcasts and multicasts).

But using bridges also has a few disadvantages:

- Bridges propagate multicast and broadcast frames. This allows broadcasts to travel throughout your network, making stations vulnerable to floods of broadcast traffic that may be generated by network software bugs, poorly designed software, or inadvertent network loops on a switch that doesn't support the Spanning Tree Protocol.
- Bridges typically cannot load-share across multiple network paths. However, you may be able to use the link aggregation protocol to provide load sharing capabilities across multiple aggregated links.

Advantages of using routers include the following:

- Routers automatically direct traffic to specific portions of the network based on the Internet Protocol (IP) destination address, providing better traffic control.
- Routers block the flow of broadcasts and multicasts. Routers also structure the flow of traffic throughout a network system based on Layer 3 network protocol addresses. This allows you to design more complex network topologies, while still retaining high stability for network operations as your network system grows and evolves.
- Routers use routing protocols that can provide information such as the bandwidth of a path. Using that information, routers can provide best-path routing and use multiple paths to provide load sharing.
- Routers provide greater network manageability in terms of access control filters and restricting access based on IP addresses.

Again, though, there are a few drawbacks to using routers:

- Router operation is not automatic, making routers more complex to configure.
- Routers may be more expensive and may provide fewer ports than switches.

The state of the art for bridges and routers is constantly evolving, and today many high-end switches are capable of operating as bridges and routers simultaneously, combining Layer 2 bridging and Layer 3 routing capabilities in the same device. You need to evaluate

these approaches to establishing a network design and building a network system, given the requirements at your site.

Special-Purpose Switches

Previously, we have described basic switch operation and features. Ethernet switches are building blocks of modern networking, and switches are used to build every kind of network that is imaginable. To meet these needs, vendors have created a wide range of Ethernet switch types and switch features. In this section, we cover several special-purpose switches, many developed for specific network types. The Ethernet switch market is a big place, though, and here we can only provide an overview of some of the different kinds of switches that are available for specific markets. There are switches designed for enterprise and campus networks, data center networks, Internet service provider (ISP) networks, industrial networks, and more. Within each category there are also multiple switch models.

Multilayer Switches

As networks became more complex and switches evolved, the development of the multilayer switch combined the roles of bridging and routing in a single device. This made it possible to purchase a single switch that could perform both kinds of packet forwarding: bridging at Layer 2 and routing at Layer 3. Early bridges and routers were separate devices, each with a specific role to play in building networks. Ethernet switches typically provided high-performance bridging across a lot of ports, and routers specialized in providing high-level protocol forwarding (routing) across a smaller set of ports. By combining those functions, a multilayer switch could provide benefits to the network designer.

As you might expect, a multilayer switch is more complex to configure than either a dedicated bridge or a router. However, by providing both bridging and routing functionality in the same device, a multilayer switch makes it easier to build large and complex networks that combine the advantages of both forms of packet forwarding. This makes it possible for vendors to provide high-performance operation of both bridging and routing across a large set of Ethernet ports at a competitive price point.

Large multilayer switches are often used in the core of a network system, to provide both Layer 2 switching and Layer 3 routing, depending on requirements. As networks grow, Layer 3 routing can provide isolation between Ethernet systems and help enable a network plan based on a hierarchical design. Multilayer switches are also used as distribution switches in building networks, providing an aggregation point for access switches. Layer 3 routing in an aggregation switch can provide separate Layer 3 networks per VLAN in a building, improving isolation, resiliency, and performance.

Access Switches

In a large enterprise network, the bulk of the network connections are on the edge, where access switches are used to connect to end nodes such as desktop computers. As a result, there is a large marketplace for access switches, with multiple vendors and a wide variety of features and price points.

When it comes to building large networks with dozens or even hundreds or thousands of access switches, a major consideration is the set of features that can provide ease of monitoring and management. Other considerations may include whether the access switches support high-speed uplinks, features like multicast packet management, and internal switching speeds capable of handling all ports running at maximum packet rates.

Each vendor has a story to tell about the capabilities of its access switches and how they can help make it easier for you to build and manage a network. While it's a significant task to compare and contrast the various access switches, you can learn a lot in the process, and it will help you make an informed purchasing decision.

Stacking Switches

Some switches are designed to allow "stacking," or combining a set of switches to operate as a single switch. Stacking makes it possible to combine multiple switches that support 24 or 48 ports each, for example, and manage the stacked set of switches and switch ports as a single logical switch that supports the combined set of ports. Stacking also provides benefits when replacing a failed switch that is a member of the stack, because the replacement switch can be automatically reconfigured with the software and configuration from the failed switch, making it possible to quickly and easily restore network service.

Stacking switches are linked together with special cables to create the physical and logical stack, as shown in Figure 19-7. These cables are typically kept as short as possible, and the switches are placed directly on top of each other, forming a compact stack of equipment that functions as a single switch. There is no IEEE standard for stacking; each vendor offering this feature has its own stacking cables and connectors.

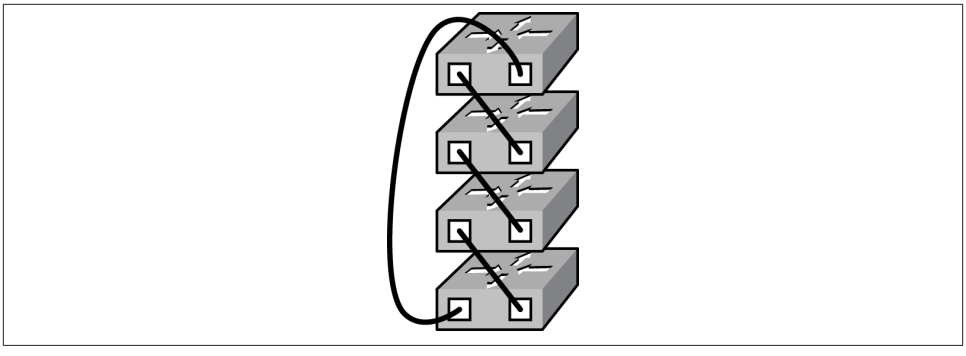


Figure 19-7. Stacking switches

The stacking cables create a switching backplane between the switches, but the packet switching speeds supported between switches in a stacking system vary between vendors. Some stacking switches are designed to use 10 Gigabit Ethernet cables connected between their uplink ports to create a stack. In this case, the stacking cable is a standard Ethernet patch cable, but the stacking software running on the switches is specific to the vendor. As a result, you cannot mix and match stacking switches from different vendors.

Industrial Ethernet Switches

Industrial Ethernet switches are switches that have been "hardened" to make it possible for them to function in the harsh environments found in factories and other industrial settings. These switches are used to support industrial automation and control systems, as well as network connections to instrumentation for the control and monitoring of major infrastructures such as the electrical power grid.

Industrial Ethernet switches may also feature special port connections that provide a seal around the Ethernet cable to keep moisture and dirt out of the switch ports. The switch itself may be sealed and fanless, to avoid exposing the internal electronics to a harsh environment. You may also find ratings for the G-forces and vibration that industrial switches can handle. To make it possible to meet the stringent environmental specifications, industrial switches are often built as smaller units, with a limited set of ports.

Wireless Access Point Switches

The development of wireless Ethernet switches is a recent innovation. "Wireless Ethernet" is a marketing term for the IEEE 802.11 wireless LAN (WLAN) standard, which provides wireless data services, also known as "Wi-Fi." Wireless LAN vendors provide a wide range of approaches to building and running a WLAN, with one widely adopted approach consisting of access points (APs) connected to WLAN controllers. A wireless

LAN controller provides support for the APs, helping to maintain the operation of the wireless system through such functions as automatic wireless channel assignment and radio power management. Typically, the controller is located in the core of a network, and wireless AP data flows are connected to the controller either directly over a VLAN, or via packet tunneling over a Layer 3 network system.

Wireless vendors have extended the controller approach by combining the controller software with a standard Ethernet switch. Including wireless controller functions and wired Ethernet ports in a single device reduces cost and makes it possible for the wireless system to grow larger without overloading a central controller. Wireless switches typically include ports equipped with Power over Ethernet, powering the APs and managing their data and operation in a single device.

A typical wireless user in an enterprise or corporate network is required to authenticate himself before being allowed to access the wireless system, which improves security and makes it possible to manage the system down to the level of the individual user. Wireless Ethernet switches can also provide client authentication services on wired ports, extending the wireless user management capabilities to the wired Ethernet system.

Internet Service Provider Switches

In recent years, Ethernet has made inroads in the Internet service provider marketplace, providing switches that are used by carriers to build wide area networks that span entire continents and the globe. ISPs have specialized requirements based on their need to provide multiple kinds of networking. ISPs provide high-speed and high-performance links over long distances, but they may also provide metro-area links between businesses in a given city, as well as links to individual users to provide Internet services at homes.

None of the ISP and carrier networks are particularly "local" in the sense of the original "local area network" design for Ethernet. The development of full-duplex Ethernet links freed Ethernet from the timing constraints of the old half-duplex media system, and made it possible to use Ethernet technology to make network connections over much longer distances than envisioned in the original LAN standard. With the local area network distance limitations eliminated, carriers and ISPs were then able to exploit the high-volume Ethernet market to help lower their costs by providing Ethernet links for long distances, metro areas, and home networks.

Metro Ethernet

While the IEEE standardizes Ethernet technology and basic switch functions, there are sets of switch features that are used in the carrier and ISP marketplaces to meet the needs of those specific markets. To meet these requirements, businesses have formed forums and alliances to help identify the switch features that are most important to them, and to publish specifications that include specialized methods of configuring switches that improve interoperability between equipment from different vendors.

One example of this is the Metro Ethernet Forum (MEF).[4] In January 2013, the MEF announced the certifications for equipment that can be configured to meet Carrier Ethernet 2.0 (CE 2.0) specifications. The set of CE 2.0 specifications include Ethernet switch features that provide a common platform for delivering services of use to carriers and ISPs. By carefully specifying these services, the MEF seeks to create a standard set of services and a network design that is distinguished from typical enterprise Ethernet systems by five "attributes." The attributes include services of interest to carriers and ISPs, along with specifications to help achieve scalability, reliability, quality of service, and service management.

The Metro Ethernet Forum is an example of a business alliance developing a comprehensive set of specifications that help define Ethernet switch operation for metro, carrier, and ISP networks. These specifications rely on a set of switch features, some of them quite complex, that are needed to achieve their network design goals, but which typically don't apply to an enterprise or campus network design.

Data Center Switches

Data center networks have a set of requirements that reflect the importance of data centers as sites that host hundreds or thousands of servers providing Internet and application services to clients. Corporate data centers hold critically important servers whose failure could affect the operations of major portions of the company, as well as the company's customers. Because of the intense network performance requirements caused by placing many critically important servers in the same room, data center networks use some of the most advanced Ethernet switches and switch features available.

Data center port speeds

Some data center servers provide database access or storage access used by other high-performance servers in the data center, requiring 10 Gb/s ports to avoid bottlenecks. Major servers providing public services, such as access to web pages, may also need 10 Gb/s interfaces, depending on how many clients they must serve simultaneously.

Data center servers also host virtual machines (VMs) in which a single physical server runs software that functions as multiple virtual servers. There can be many VMs per server, and hundreds or thousands of VMs per data center. As each VM is a separate operating system running separate services, the large number of VMs also increases the network traffic generated by each physical server.

4. According to the MEF website, it is "a global industry alliance comprising more than 200 organizations including telecommunications service providers, cable MSOs, network equipment/software manufacturers, semiconductors vendors and testing organizations. The MEF's mission is to accelerate the worldwide adoption of Carrier-class Ethernet networks and services. The MEF develops Carrier Ethernet technical specifications and implementation agreements to promote interoperability and deployment of Carrier Ethernet worldwide."

Data center switches typically feature non-blocking internal packet switching fabrics, to avoid any bottlenecks internal to the switches. These switches also feature high-speed ports, because modern servers are equipped with Gigabit Ethernet interfaces and many of them come with interfaces that can run at both 1 and 10 Gb/s.

Data center switch types

Data centers consist of equipment cabinets or racks aligned in rows, with servers mounted in the racks in dense stacks, using mounting hardware that screws into flanges located in the racks. When you open the door to one of these cabinets, you are faced with a tightly packed stack of servers, one on top of the other, filling the cabinet space. One method for providing switch ports in a cabinet filled with servers is to locate a *top of rack* (TOR), switch at the top of the cabinet and to connect the servers to ports on that switch.

The TOR switch connects, in turn, to larger and more powerful switches located either in the middle of a row, or in the end cabinet of a row. The row switches, in their turn, are connected to core switches located in a core row in the data center. This provides a "core, aggregation, edge" design. Each of the three types of switches has a different set of capabilities that reflect the role of the switch.

The TOR switch is designed to be as low-cost as possible, as befits an edge switch that connects to all stations. However, data center networks require high performance, so the TOR switch must be able to handle the throughput. The row switches must also be high-performance, to allow them to aggregate the uplink connections from the TOR switches. Finally, the core switches must be very high-performance and provide enough high-speed uplink ports to connect to all row switches.

Data center oversubscription

Oversubscription is common in engineering: it describes a system that is provisioned to meet the typical demand on the system rather than the absolute maximum demand. This reduces expense and avoids the purchasing of resources that would rarely, or never, be used. In network designs, oversubscription makes it possible to avoid purchasing excess switches and paying for higher port performance.

When you're dealing with the hundreds or thousands of high-performance ports that may be present in a modern data center, it can be quite difficult to provide enough bandwidth for the entire network to be "non-blocking," meaning that all ports can simultaneously operate at their maximum performance. Rather than provide a non-blocking system, a network that is oversubscribed can serve all users economically, without a significant impact on their performance.

As an example, a non-blocking network design for 100 10 Gb/s ports located on a single row of a data center would have to provide 1 terabit (Tb) of bandwidth to the core switches. If all 8 rows of a data center each needed to support 100 10 Gb/s ports, that

would require 8 Tb of port speed up to the core, and an equivalent amount of switching capability in the core switches. Providing that kind of performance is very expensive.

And even if you had the money and space to provide that many high-performance switches and ports, the large investment in network performance would be wasted, because the bandwidth would be mostly unused. Although a given server or set of servers may be high-performance, in the vast majority of data centers not all servers are running at maximum performance all of the time. In most network systems, the Ethernet links tend to run at low average bit rates, interspersed with higher-rate bursts of traffic. Thus, you do not need to provide 100% simultaneous throughput for all ports in the majority of network designs, including most data centers. Instead, typical data center designs can include a significant level of oversubscription without affecting performance in any significant way.

Data center switch fabrics

Data centers continue to grow, and server connection speeds continue to increase, placing more pressure on data center networks to handle large amounts of traffic without excessive costs. To meet these needs, vendors are developing new switch designs, generally called "data center switch fabrics." These fabrics combine switches in ways that increase performance and reduce oversubscription.

Each of the major vendors has a different approach to the problem, and there is no one definition for a data center fabric. There are both vendor-proprietary approaches and standards-based approaches that are called "Ethernet fabrics," and it is up to the marketplace to decide which approach or set of approaches will be widely adopted. Data center networks are evolving rapidly, and network designers must work especially hard at understanding the options and their capabilities and costs.

Data center switch resiliency

A major goal of data centers is high availability because any failure of access can affect large numbers of people. To achieve that goal, data centers implement resilient network designs based on the use of multiple switches supporting multiple connections to servers. As with other areas of network design, resilient approaches are evolving through the efforts of multiple vendors and the development of new standards.

One way to achieve resiliency for server connections in a data center is to provide two switches in a given row, call them Switch A and Switch B, and to connect the server to both switches. To exploit this resiliency, some vendors provide multichassis link aggregation (MLAG), in which software running on both switches makes the switches appear as a single switch to the server. The server thinks that it is connected to two Ethernet links on a single switch using standard 802.1AX link aggregation (LAG) protocols. But in reality, there are two switches involved in providing the aggregated link (hence the name multichassis LAG). Should a port or an entire switch fail for any reason, the server

still has an active connection to the data center network and will not be isolated from the network.

Advanced Switch Features

While the common features found on most switches are sufficient for the needs of most networks, switches designed for specific kinds of networks may provide extra features that are specific to the networks involved. In this section, we describe advanced features that may be found in a variety of switches, as well as specialized features found in switches designed for specific networking environments.

Traffic Flow Monitoring

Given that they are providing the infrastructure for switching packets, switches can provide useful management data on traffic flows through your network. By collecting data from multiple switches, or by collecting data at the core switches, you can be provided with views of network traffic that are valuable for monitoring network performance and predicting the growth of traffic and the need for more capacity in your network.

As usual in the networking industry, there are multiple standards and methods for collecting data from switches. We've already mentioned one widely used system called the Simple Network Management Protocol (see Chapter 18), which can be used to collect packet counts on ports, and other operational information. However, while counting packets is useful and can help you produce valuable traffic graphs, sometimes you want further information on the traffic flowing through your network.

sFlow and NetFlow

There have been two systems developed to provide information on traffic flows, called *sFlow* and *NetFlow*. sFlow is a freely licensed specification for collecting traffic flow information from switches. NetFlow is a protocol developed by Cisco Systems for collecting traffic flow information. The NetFlow protocol has evolved to become the Internet Protocol Flow Information Export (IPFIX) protocol, which is being developed as an Internet Engineering Task Force (IETF) protocol standard.

Assuming that your switch supports sFlow, NetFlow, or IPFIX, you can collect data on network traffic flows that can provide visibility into the traffic patterns on your network. The data provided by these protocols can also be used to alert you to unusual traffic flows, including attack traffic that might otherwise not be visible to you.

If your switch does not support traffic flow software, there are still some options available. There are a number of devices that can provide sFlow and NetFlow data, using traffic exported from your switch to dedicated computers running software that can

turn the packet flows into flow records and then analyze and display the information in those records.

One method to provide flow data is to "tap" the flows of traffic on the core switch and send it to an outboard computer running packet flow software. If your switch supports packet mirroring without affecting switch performance, you could mirror the traffic onto a port and connect that port to the outboard flow analysis computer. If your main network connections are based on fiber optic Ethernet, then another method is to use fiber optic splitters to send a copy of the optical data to an outboard computer for analysis.

Power over Ethernet

As we saw in Chapter 6, Power over Ethernet (PoE) is a standard that provides direct current (DC) electrical power over Ethernet twisted-pair cabling, to operate Ethernet devices at the other end of the cable. For devices with relatively low power requirements, such as wireless access points, VoIP telephones, video cameras, and monitoring devices, PoE can reduce costs by avoiding the need to provide a separate electrical circuit for each device. Switch ports can be equipped to provide PoE, turning a switch into a power management point for network devices.

Many access points, telephones, and video cameras can be powered over the original PoE system, delivering up to 15.4 watts of DC current over the Ethernet cable. However, some devices—such as newer access points with more electronics, or video cameras with motors for zoom, pan, and tilt functions—may draw more wattage. The revision of the PoE standard developed as part of the 820.3at supplement in 2009 provides up to 30 watts, and some vendors have gone beyond the standard and are providing even higher amounts (up to 60 watts), by sending the DC current over all four pairs of a twisted-pair cable.

While power can be injected into Ethernet cables with an outboard device, a convenient method is to use the switch port as the power sourcing equipment (PSE). A standard PSE provides approximately 48 volts of direct current power to the powered device (PD) over two pairs of twisted-pair cabling. There is also a management protocol that makes it possible for the PD to inform the PSE about its requirements, allowing the PSE to avoid sending unnecessary power over the cable.

With multiple switch ports acting as PSEs, there can be a significant increase in the amount of power required by a given switch. If you plan to use a single switch to provide PoE to many devices, you need to investigate the total power requirements, to make sure that the power supply on the switch can handle the load and that the electrical circuit that the switch uses is able to provide the amount of current required.

PART V
Performance and Troubleshooting

This part will discuss important topics in Ethernet performance and troubleshooting. Chapter 20 describes both the performance of a given Ethernet channel, and the performance of the network system as a whole. Chapter 21 includes a tutorial on troubleshooting techniques and describes the kinds of problems you are likely to encounter with twisted-pair and fiber optic systems.

CHAPTER 20
Ethernet Performance

"Performance" is an umbrella term that can mean different things to different people. To a network designer, the performance of an Ethernet system can range from the performance of individual Ethernet channels, to the performance of Ethernet switches, to the performance and capabilities of the entire network system.

For the users of a network, on the other hand, performance usually refers to how quickly the applications that they are using over the network respond to their commands. In this case, the performance of the Ethernet system that connects to a user's computer is only one component in a whole set of entities that must work together to provide a good user experience.

Because this is a book about Ethernet local area networks, we will focus on the performance of the Ethernet channel and the network system. Along the way, we will also show how the performance of the network is affected by a complex set of elements that includes local servers, filesystems, cloud servers, the Internet, and the local Ethernet system, all working to provide application services for users.

The first part of this chapter discusses the performance of the Ethernet channel itself. We will examine some of the theoretical and experimental analytical techniques that have been used to determine the performance of a single Ethernet channel. Later, we discuss what reasonable traffic levels on a real-world Ethernet can look like. We also describe what kinds of traffic measurements you can make, and how to make them.

In the last part of the chapter, we show that various kinds of traffic have different response time requirements. We also show that response time performance for the user is the complex sum of the response times of the entire set of elements used to deliver application services between computers. Finally, we provide some guidelines for designing a network to achieve the best performance.

Performance of an Ethernet Channel

The performance of an individual Ethernet channel was a major topic when all Ethernet systems operated in half-duplex mode, using the CSMA/CD media access control mechanism. On a half-duplex system, all stations used the CSMA/CD algorithm to share access to a single Ethernet channel, and the performance of CSMA/CD under load was of considerable interest.

Those days are gone, and now virtually all Ethernet channels are automatically configured by the Auto-Negotiation protocol to operate in full-duplex mode, between a station and a switch port. In full-duplex mode a particular station "owns" the channel, because the Ethernet link is dedicated to supporting a single station's connection to a switch port. There are two signal paths, one for each end of the segment, so the station and the switch port can send data whenever they like, without having to wait.

Both the station and the switch port can send data at the same time, which means that a full-duplex segment can provide twice the rated bandwidth at maximum load. In other words, a 1 Gb/s full-duplex link can provide 2 Gb/s of bandwidth, assuming that both the station and the switch port are sending the full rate of traffic in both directions simultaneously.

Performance of Half-Duplex Ethernet Channels

Ethernet systems operated as half-duplex systems for many years, and a number of simulations and analyses of those systems were published. This activity resulted in a number of myths and misconceptions about the performance of older Ethernet systems, which are discussed next. Keep in mind that we are discussing the old half-duplex model of Ethernet operation, which is rarely used anymore.

When calculating the performance of a half-duplex Ethernet channel, researchers often used simulations and analytic models based on a deliberately overloaded system. This was done to see what the limits of the channel were, and how well the channel could hold up to extreme loads.

Over time, the simulations and analytic models used in these studies became increasingly sophisticated in their ability to model actual half-duplex Ethernet channel behavior. Early simulations frequently made a variety of simplifying assumptions to make the analysis easier, and ended up analyzing systems whose behavior had nothing much to do with the way a real half-duplex Ethernet functioned. This produced some odd results, and led some people to deduce that Ethernets would saturate at low utilization levels.

Persistent Myths About Half-Duplex Ethernet Performance

Due to the incorrect results coming from simplified models, there arose some persistent myths about half-duplex Ethernet performance, chief of which was that the Ethernet

channel saturated at 37% utilization. We'll begin with a look at where this figure comes from, and why it had nothing to do with real-world Ethernets.

The 37% figure was first reported by Bob Metcalfe and David Boggs in their 1976 paper that described the development and operation of the very first Ethernet.[1] This was known as the "experimental Ethernet," which operated at about 3 Mb/s. The experimental Ethernet frame had 8-bit address fields, a 1-bit preamble, and a 16-bit CRC field.

In this paper, Metcalfe and Boggs presented a "simple model" of performance. Their model used the smallest frame size and assumed a constantly transmitting set of 256 stations, which was the maximum supported on experimental Ethernet. Using this simple model, the system reached saturation at about 36.8% channel utilization. The authors warned that this was a simplified model of a constantly overloaded shared channel, and did not bear any relationship to normally functioning networks. However, this and subsequent studies based on the simplified model as applied to 10 Mb/s Ethernet led to a persistent myth that "Ethernet saturates at 37% load."

This myth about Ethernet performance persisted for years, probably because no one understood that it was merely a rough measure of what could happen if one used a very simplified model of Ethernet operation and absolute worst-case traffic load assumptions. Another possible reason for the persistence of this myth was that this low figure of performance was used by salespeople in an attempt to convince customers to buy competing brands of network technology and not Ethernet.

In any event, after years of hearing people repeat the 37% figure, David Boggs and two other researchers published a paper in 1988, entitled "Measured Capacity of an Ethernet: Myths and Reality."[2] The objective of this paper was to provide measurements of a real Ethernet system that was being pushed very hard, which would serve as a corrective for the data generated by theoretical analysis that had been published in the past.

The three authors of the paper, Boggs, Mogul, and Kent, noted that these experiments did not demonstrate how an Ethernet normally functions. Ethernet, in common with many other LAN technologies, was initially designed to support "bursty" traffic instead of a constant high traffic load. In normal operation, many half-duplex Ethernet channels operated at fairly low loads, averaged over a five-minute period during the business day, interrupted by peak traffic bursts as stations happened to send traffic at approximately the same time.

1. Robert M. Metcalfe and David R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," *Communications of the ACM* 19:5 (July 1976): 395–404.

2. David R. Boggs, Jeffrey C. Mogul, and Christopher A. Kent, "Measured Capacity of an Ethernet: Myths and Reality,"*Proceedings of the SIGCOMM '88 Symposium on Communications Architectures and Protocols*, (August 1988), 222–234.

In the Boggs, Mogul, and Kent paper, a population of 24 workstations were programmed to constantly flood a 10 Mb/s Ethernet channel in several experimental trials, each using a different frame size, and some using mixed frame sizes. The results showed that the Ethernet channel was capable of delivering data at very high rates of channel utilization even when the 24 stations were constantly contending for access to the channel. For small frames sent between a few stations, channel utilization was as high as 9 Mb/s, and for large frames utilization was close to the maximum of 10 Mb/s (100% utilization).

With 24 stations running full blast, there was no arbitrary saturation point of 37% utilization (confirming what LAN managers who had been using Ethernet for years already knew). Nor did the system collapse with a number of stations offering a high constant load, which had been another popular myth. Instead, the experiments demonstrated that a half-duplex Ethernet channel could transport high loads of traffic among this set of stations in a stable fashion and without major problems.

Figure 20-1 shows a graph of Ethernet utilization from the Boggs, Mogul, and Kent paper[3] that illustrates the maximum channel utilization achieved when up to 24 stations were sending frames continuously, using a variety of frame sizes. The frame sizes for each graph are numbered from 1 through 10, and range from 64 bytes (graph number 10) on up to 4,000 bytes (graph number 1). Any frame larger than 1,518 bytes exceeds the maximum allowed in the Ethernet specification, but the larger frame sizes were included in this test to see what would happen when the channel was stress-tested in this way. The graph shows that even when 24 stations were constantly contending for access to the channel, and all stations were sending small (64-byte) frames, channel utilization stayed quite high, at around 9 Mb/s.

The Boggs, Mogul, and Kent paper also provides some guidelines for network design based on their analysis. Two points they made are worth restating here:

- Don't put too many stations on a single half-duplex channel (and therefore in a single collision domain). For best performance, use switches and routers to segment the network into multiple Ethernet segments.
- Avoid mixing heavy use of real-time applications with bulk-data applications. High traffic loads on the network caused by bulk-data applications produce higher transmission delays, which will negatively affect the performance of real-time applications. (We will discuss this issue in more detail later in this chapter.)

3. Boggs, Mogul, and Kent, "Measured Capacity of an Ethernet," Figure I-1, p. 24, used by permission.

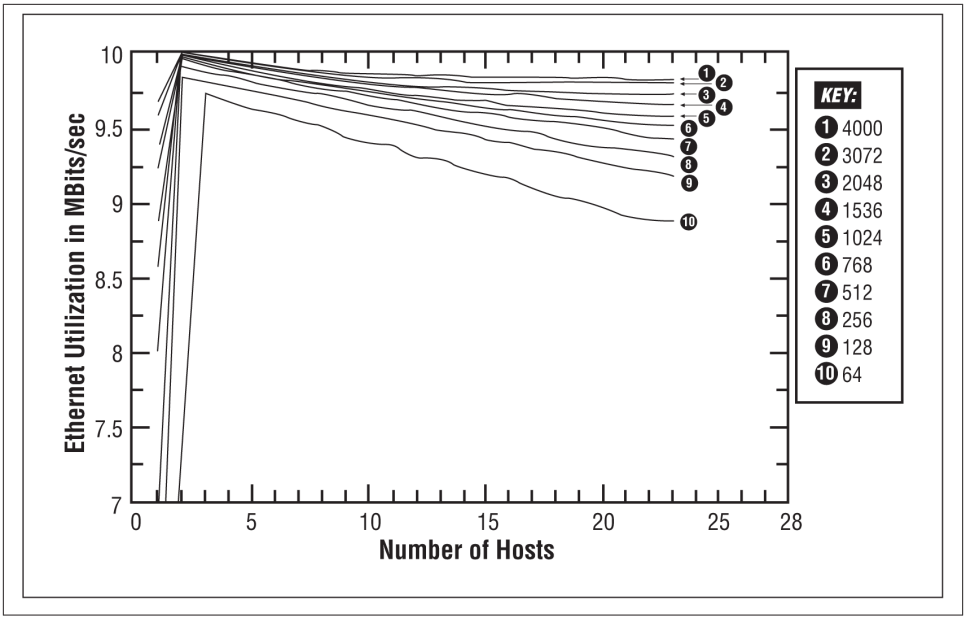


Figure 20-1. Ethernet utilization graph

Simulations of Half-Duplex Ethernet Channel Performance

The Boggs, Mogul, and Kent paper noted that some of the theoretical studies that had been made of Ethernet performance were based on simulations that did not appear to accurately model Ethernet behavior. Building an accurate simulator of Ethernet behavior is difficult, because transmissions on an Ethernet are not centrally controlled in any way; instead, they happen more or less randomly, as do collisions.

In 1992, Speros Armyros published a paper showing the results of a new simulator for Ethernet that could accurately duplicate the real-world results reported in the Boggs, Mogul, and Kent paper.[4] This simulator made it possible to try out some more stress tests of the Ethernet system.

These new tests replicated the results of the Boggs, Mogul and Kent paper for 24 stations. They also showed that under worst-case overload conditions, a single Ethernet channel with over 200 stations continually sending data would behave rather poorly, and access times would rapidly increase. *Access time* is the time it takes for a station to transmit a packet onto the channel, including any delays caused by collisions and by multiple packets backing up in the station's buffers due to congestion of the channel.

4. Speros Armyros, "On the Behavior of Ethernet: Are Existing Analytic Models Accurate?" Technical Report CSRI-259, February 1992, Computer Systems Research Institute, University of Toronto, Toronto, Canada.

Further analysis of an Ethernet channel using the improved simulator was published by Mart Molle in 1994.[5] Molle's analysis showed that the Ethernet *binary exponential backoff* (BEB) algorithm was stable under conditions of constant overload on Ethernet channels with station populations under 200. However, once the set of stations increased much beyond 200, the BEB algorithm began to respond poorly. In this situation (under conditions of constant overload), the access time delays encountered when sending packets can become unpredictable, with some packets encountering rather large delays.

Molle also noted that the capture effect, described in Appendix B, can actually improve the performance of an Ethernet channel for short bursts of small packets. However, the capture effect also leads to widely varying response times when trains of long packets briefly capture the channel. Finally, Molle's paper described a new backoff algorithm that he created to resolve these and other problems, called the *Binary Logarithmic Arbitration Method* (BLAM). BLAM was never formally adopted by the Ethernet standard, for the reasons explained in Appendix B.

Molle noted that constantly overloaded channels are not a realistic model of real-world usage. What the network users are interested in is response time, which includes the typical delay encountered when transmitting packets. Highly congested channels exhibit very poor response times, which users find unacceptable.

Figure 20-2 shows a graph from Molle's paper, displaying the effects that channel load and the number of stations (hosts) have on response time.[6] The chart shows that the average channel response time is good until the channel is seeing a constant load of more than 50%. The region from 50% constant load to about 80% constant load shows increased delays, and above 80% constant load the delays increase rapidly.

5. Mart M. Molle, "A New Binary Logarithmic Arbitration Method for Ethernet," Technical Report CSRI-298, April 1994 (revised July 1994), Computer Systems Research Institute, University of Toronto, Toronto, Canada.

6. Molle, "A New Binary Logarithmic Arbitration Method for Ethernet," Figure 5, p. 13, used by permission.

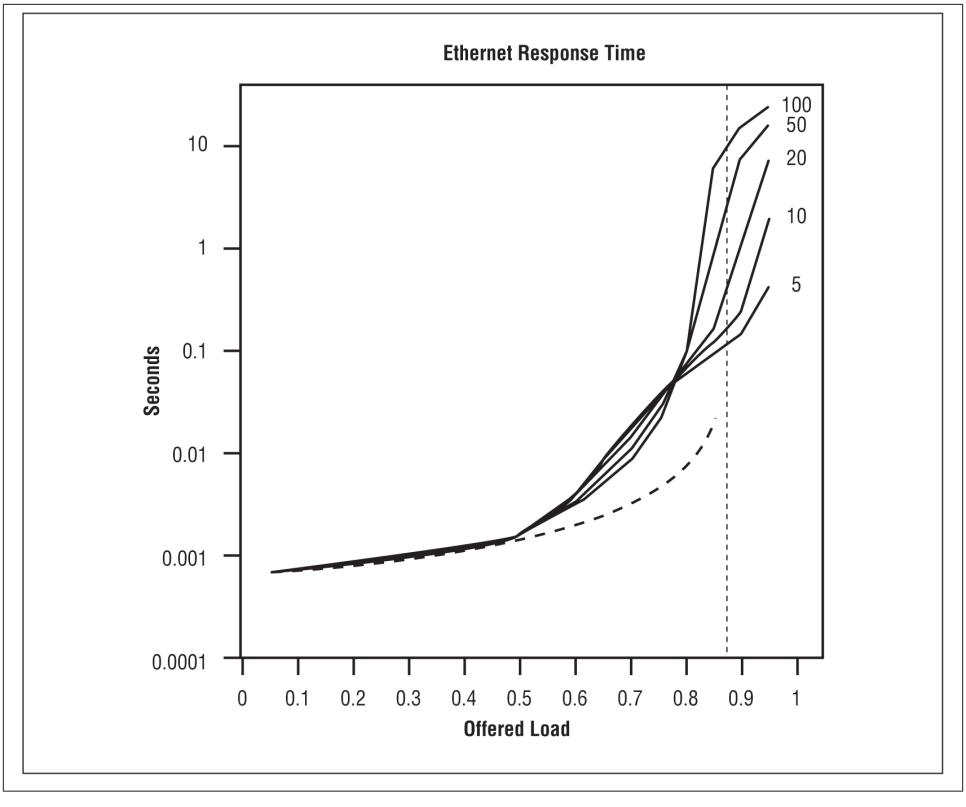


Figure 20-2. Ethernet channel access time

Another element is the variation in channel access times, known as *jitter*. For example, real-time traffic carrying audio information will work best on an uncongested channel that provides rapid response times. Heavily loaded channels can result in excessive delay and jitter, making the audio bursty and difficult to understand. Therefore, excessive jitter will be unacceptable to the users of real-time applications.

The Molle study shows that there are three major operating regimes for an Ethernet half-duplex channel:

Light load

Up to 50% average utilization measured over a one-second sample time. At this level of utilization, the network responds rapidly. Stations can send packets with very low access delays of about 0.001 seconds or less on a 10 Mb/s channel. The response of real-time applications will be acceptable.

Moderate to heavy load
: From 50 to 80% average utilization measured over a one-second sample period. At this point, the channel begins to show larger delays in the range of 0.01 to 0.1 seconds on a 10 Mb/s channel.

  This kind of transmission delay will not be noticeable to applications such as web browsers, or when accessing file servers or databases. However, transmission delays may be large enough for some packets that real-time applications could experience negative effects from variable delay. Short-term traffic bursts into this region should not be a problem, but longer-term load averages at this rate of utilization are not recommended for best performance.

Very high load
: From 80 to 100% average utilization measured over a one-second sample period. At this rate, the transmission delays can get quite high, and the amount of jitter can get very large. Access delays of up to a second are possible on 10 Mb/s channels, while even longer delays have been predicted in simulations. Short-term traffic bursts into this region should not be a problem, but long-term average loads at this rate would indicate a seriously overloaded channel.

The lessons learned in these studies make it clear that users trying to get work done on a constantly overloaded channel will perceive unacceptable delays in their network service. Although constant high network loads may feel like a network "collapse" to the users, the network system itself is in fact still working as designed; it's just that the load is too high to rapidly accommodate all of the people who wish to use the network.

The delays caused by congestion in communications channels are somewhat like the delays caused by congested highways during commute times. Despite the fact that you are forced to endure long delays due to a traffic overload of commuters, the highway system isn't broken; it's just too busy. An overloaded Ethernet channel has the same problem.

Measuring Ethernet Performance

Now that we've seen the analysis of half-duplex Ethernet channels, let's look more closely at how to monitor a normally operating modern Ethernet system, where virtually all channels are operating in full-duplex mode. In this mode, the channel can be loaded to 100% without affecting the channel access time for a station. That's because full-duplex channels provide dedicated signal paths for the devices at each end of the channel, allowing them to send data whenever they like and allowing the channel to operate at loads up to 100% without affecting access to the channel.

Monitoring the total amount of traffic on a given Ethernet link requires a device that operates in *promiscuous receive mode*, reading in every frame seen on the channel.

Looking at every frame with a general-purpose computer requires a network interface and computer system that can keep up with high frame rates.

In the older Ethernet systems based on coaxial cables and shared channels, you could attach a monitor to the cable and see all of the traffic from all stations on that channel. These days it's considerably harder to monitor an Ethernet system, because Ethernets are built using individual Ethernet segments connected to switch ports.

Therefore, you need to monitor the switch itself, or use a monitoring device connected to a packet mirror or SPAN port on the switch. Higher-cost switches also have built-in management, which allows you to monitor the utilization and other statistics on each port as well as for the entire device. One useful method for collecting utilization and other statistics is provided by the Simple Network Management Protocol (SNMP). Using SNMP-based management software is discussed in more detail in Chapter 21.

Measurement Time Scale

Before you set out to measure the network load on a real network, you need to determine what time scale to use. Many network analyzers are set by default to look at the load on an Ethernet averaged over a period of one second. One second may not sound like a long period of time, but a 10 Mb/s Ethernet channel operating flat out can theoretically transmit 14,880 frames in that one second. A Gigabit Ethernet system can send a hundred times that amount, or 1,488,000 frames per second, and a 100 Gigabit system can send 148,800,000 frames per second.

By looking at the load on a network in one-second increments, you can generate a set of data points that can be used to draw a graph of the one-second average loads over time. This graph will rise and fall, depending on the average network traffic seen during the one-second sample period.

The one-second sample time can be useful when looking at the performance of a network port in real time. However, most network systems consist of more than one network segment, and most network managers have better things to do than spend all day watching the traffic loads on Ethernet ports. Management software exists that will automatically create reports of network utilization and store them in a database. These reports are used to create charts that show traffic over the busy hours of a workday, or the entire day, week, or month.

Because the traffic rate on a LAN can vary significantly over time, you really need to look at the average utilization over several time periods to get an idea of the general loads seen on the network. The traffic on most Ethernet ports tends to be bursty, with large, short-lived peaks. Peak network loads can easily go to 80, 90, or 100% measured over a one-second interval without causing any problems for a typical mix of applications.

For an Ethernet system, the minimal set of things you might consider keeping track of include:

- The utilization rate of the network ports over a series of time scales.
- The rate of broadcasts and multicasts. Excessive rates of broadcasts can affect station performance because every station must read in every broadcast frame and decide what to do with it.
- Basic error statistics, including cyclic redundancy check (CRC) errors, oversize frames, and so on.

The time scales you choose for generating utilization figures are a matter of debate, because no two networks are alike and every site has a different mix of applications and users. Network managers often choose to create baselines of traffic that extend over several time scales. With the baselines stored, they can then compare new daily reports against previous reports to make sure that the Ethernet ports are not staying at high loads during the times that are important to the users.

Most networks are designed in a hierarchy, with access switches connected to core switches using uplink ports. Monitoring those uplink ports is critical, because they are the bottlenecks for your network system. Very high loads on those uplinks that last for significant periods of time during the workday could cause unacceptable delays.

For example, filesystem backups can take a fair amount of time to perform and often place a heavy demand on your network. To ensure that your users have priority access to the network, it's recommended that you perform backups at night, when the likelihood of user traffic on the network is minimal.

Constant monitoring also provides evidence of overload that can be useful when responding to complaints about network performance. Given the wide variability in application mix, number of users, and so on, it is quite difficult to provide any rules of thumb when it comes to network load. Some network managers report that they regard network traffic as approaching excessive load levels when:

- Uplink utilization averaged over the eight-hour workday exceeds 20%.
- Average utilization during the busiest hour of the day exceeds 30%.
- Fifteen-minute averages exceed 50% at any time during the workday.

Notice that these recommendations are not based on the three operating regimes derived from Molle's paper. The three operating regimes that Molle studied are based on one-second average loads on half-duplex shared channels. In the traffic load levels just shown, an eight-hour average utilization that reaches 20% is a heavily smoothed graph, which does not show the short-term peaks. During the business day, we can assume that transient peaks went much higher than 20%. More importantly, we can assume that

when the long-term average gets that high, the peak traffic loads may have been lasting for long periods, producing unacceptable response times for the users.

There are many ways to generate graphs and reports of network utilization. Table 20-1 displays some raw data collected with an SNMP-based management program. These samples were collected every 30 minutes for the total number of packets, octets, broadcasts, and multicasts seen on an uplink port.

Table 20-1. SNMP data output

Timestamp	Packets	Octets	Broadcast	Multicast	Utilization
09:42:10	138243	41326186	882	383	2
10:12:10	161295	51701901	828	397	2
10:42:10	168389	58580988	868	391	3
11:12:10	2775468	559286267	1283	280	25
11:42:10	604774	111504337	1231	275	5
12:12:10	836423	126693664	1218	415	6
12:42:10	164848	59062247	1117	500	3
13:12:10	221535	94692849	1343	980	4

The average utilization on the channel over the 30-minute period is also collected. Notice that during the 30-minute period from 10:42 to the next sample at 11:12, the average utilization was 25%. This average is high for such a long period of time in the middle of the workday, and network users may have complained about poor response time during this period. A shorter sample time would very likely have shown much higher peak loads lasting for significant periods of time, which could cause poor response times and generate complaints about network performance.

When collecting utilization information, it's up to you to determine what load levels are acceptable to your users, given the application mix at your site. Note that short-term averages may reach 100% load for a few seconds without generating complaints. Short-term peaks such as this can happen when large file transfers cause high loads for a brief period. For many applications, the users may never notice the short-term high loads. However, if the network is being used for real-time applications, then even relatively short-term loads could cause problems.

When the reports for an Ethernet begin to show a number of high utilization periods, a LAN manager might decide to keep a closer eye on the network. You want to see whether the traffic rates are stable, or if the loads are increasing to the point where they may affect the operation of the network applications being used. The network load can be adjusted by increasing the speed of the Ethernet links, especially the uplinks between switches.

Data Throughput Versus Bandwidth

The analytical studies we've seen so far were interested in measuring total channel utilization, which includes all application data being sent as well as the framing bits and other overhead it takes to send the data. This is useful if you're looking at the theoretical bandwidth of an Ethernet channel. On the other hand, most users want to know how much data they can get through the system. This is sometimes referred to as *throughput*. Note that bandwidth and throughput are different things.

Bandwidth is a measure of the capacity of a link, typically provided in bits per second (bps). The bandwidth of Ethernet channels is rated at 10 million bits per second (10 Mb/s), 100 million bits per second (100 Mb/s), 1 billion bits per second (1 Gb/s), and so on. *Throughput* is the rate at which usable data can be sent over the channel. While an Ethernet channel may operate at 10 Mb/s, the throughput in terms of usable data will be less due to the number of bits required for framing and other channel overhead.

On an Ethernet channel, it takes a certain number of bits, organized as an Ethernet frame, to carry data from one computer to the other. The Ethernet system also requires an interframe gap between frames, and a frame preamble at the front of each frame. The framing bits, interframe gap, and preamble constitute the necessary overhead required to move data over an Ethernet channel. As you might expect, the smaller the amount of data carried in the frame, the higher the percentage of overhead. Another way of saying this is that frames carrying large amounts of data are the most efficient way to transport data over the Ethernet channel.

Maximum data rates on Ethernet

We can determine the maximum data rate that a single station can achieve by using the sizes of the smallest and largest frames to compute the maximum throughput of the system. Our frame examples include the widely used type field, because frames with a type field are easiest to describe. The IEEE 802.3 frame format with 802.2 logical link control (LLC) fields will have slightly lower performance, due to the use of a few bytes of data in the data field that are required to carry the LLC information. The numbers we come up with for the 10 Mb/s channel can simply be multiplied by 10 for a 100 Mb/s Fast Ethernet system, by 100 for Gigabit Ethernet, and so on. Keep in mind that a link operating in full-duplex mode can support twice the data rates shown here, because the devices on both ends of the link can transmit simultaneously.

The first column in Table 20-2 shows the data size (in bytes) being carried in each frame and the total frame size including the overhead bits (i.e., the non-data framing fields) in parentheses. The non-data fields of the frame include 64 bits of preamble, 96 bits of source and destination address, 16 bits of type field, and 32 bits for the frame check sequence (FCS) field, which carries the CRC.

Table 20-2. Maximum frame and data rate for 10 Mb/s Ethernet

Data field size (frame size)	Maximum frames/sec	Maximum data rate (bits/sec)
46 (64)	14,880	5,475,840
64 (82)	12,254	6,274,084
128 (146)	7,530	7,710,720
256 (274)	4,251	8,706,048
512 (530)	2,272	9,306,112
1,024 (1,042)	1,177	9,641,984
1,500 (1,538)	812	9,752,925

The interframe gap on a 10 Mb/s system is 9.6 microseconds, which is equivalent to 96 bit times. Total it all up, and we get 304 bit times of overhead required for each frame transmission. With that in mind, we can now calculate, theoretically, the number of frames that could be sent for a range of data field sizes—beginning with the minimum data size of 46 bytes, and ending with the maximum of 1,500 bytes. The results are shown in the second column of the table.

The calculations provided in Table 20-2 are made using some simplifying assumptions, as they say in the simulation and analysis trade. These assumptions are that one station sends back-to-back frames endlessly at these data sizes, and that another station receives them. This is obviously not a real-world situation, but it helps us provide the theoretical maximum data throughput that can be expected of a single 10 Mb/s Ethernet channel.

At 14,880 frames per second, the Ethernet channel is at 100% load. However, Table 20-2 shows that, while operating at 100% load, a 10 Mb/s channel moving frames with only 46 bytes of data in them can deliver a maximum of 5,475,840 bits per second of data throughput in one direction, and twice that if both directions are operating at the maximum rate on a full-duplex link. This is only about 54.7% efficiency in terms of data delivery.

If 1,500 bytes of data are sent in each frame, then an Ethernet channel operating at 100% constant load could deliver 9,744,000 bits per second of usable data for applications. This is over 97% efficiency in channel utilization. These figures demonstrate that, while the bandwidth of an Ethernet channel may be 10 Mb/s, the throughput in terms of usable data sent over that channel can vary quite a bit. It all depends on the size of the data field in the frames, and the number of frames per second.

Network performance for the user

Of course, frame size and data throughput are not the entire picture either. As far as the user is concerned, network throughput and response time are affected by the whole set of elements in the path between computers that communicate with one another. All of the following can impact the user's perception of network performance:

- The performance of the high-level network protocol software running on the user's computer.
- The overhead required by the fields in high-level protocol packets that are carried in the Ethernet frames.
- The performance of the application software being used. File-sharing performance in the face of occasional dropped packets can fall drastically depending on the amount of time required by application-level timeouts and retransmissions.
- The performance of the user's computer, in terms of CPU speed, amount of random access memory (RAM), backplane bus speed, and disk I/O speed. The performance of a bulk-data-transfer operation such as file transfer is often limited by the speed of the user's disk drive. Another limit is the speed at which the computer can move data from the network interface onto the disk drive.
- The performance of the network interface installed in the user's computer. This is affected by the amount of buffer memory that the interface is equipped with, as well the speed of the interface driver software.

As you can see, there are many elements at play. The question most network managers want answered is: "What traffic levels can the network operate at and still provide adequate performance for the users?" However, it's quite clear that this is not an easy question to answer.

Some applications require very rapid response times, while others are not that delay-sensitive. The size of the packets sent by the applications makes a big difference in the throughput that they can achieve over the network channel. Further, mixing delay-sensitive and bulk-data applications may or may not work, depending on how heavily loaded the channel is.

Performance for the network manager

So how does the network manager decide what to do? There's still no substitute for common sense, familiarity with your network system, and some basic monitoring tools. There's not much point in waiting for someone to develop a magic program that understands all possible variables that affect network behavior. Such a program would have to know all the details about the computers you are using, and how well they perform. Perhaps it would also automatically analyze your application mix and load profile and call you on the telephone to report problems.

Until the magic program arrives, you can do some basic monitoring yourself. For example, you could develop some baselines for your daily traffic so that you know how things are running today. Then you can compare future reports to the baselines to see how well things are working on a day-to-day basis.

Of course, Ethernets can run without being watched very closely, and small Ethernets may not justify monitoring at all. On a small home network supporting a few stations, you probably don't care what the load is as long as things are working. The same is probably true for many small office networks. Even large Ethernet systems—spanning an entire building or set of buildings—may have to run without analysis.

If you have no budget or staff for monitoring, then you may have very little choice except to wait for user complaints and then wade in with some analysis equipment to try to figure out what is going on. Of course, this will have a severe impact on the reliability and performance of your network, but you get what you pay for.

The amount of time, money, and effort you spend on monitoring your network is entirely up to you. Small networks won't require much monitoring, beyond keeping an eye on the error stats or load lights of your network equipment. Some switches can provide management information by way of a management interface, as described in Chapter 21. This makes it easier to monitor the error counts on the switch without investing very much money in management software. Larger sites that depend on their networks for their business operations could reasonably justify the expenditure of a fair amount of resources on monitoring. There are also companies that will monitor your network devices for you, for a fee.

Network Design for Best Performance

Many network designers would like to know ahead of time exactly how much bandwidth they will need to provide, but as we've shown in this chapter, it's not that easy. Network performance is a complex subject with many variables, and it's a distinctly nontrivial task to model a network system sufficiently well that you can predict what the traffic loads will be like.

Instead, most network managers take the same approach that highway designers take, which is to provide excess capacity for use during peak times and to accommodate some amount of future growth in the number of stations being supported. The cost of Ethernet equipment is low enough that this is fairly easy to do.

Providing extra bandwidth helps to ensure that a user can move a bulk file quickly when she needs to. Extra bandwidth also helps ensure that delay-sensitive applications will work acceptably well. In addition, once a network is installed, it attracts more computers and applications like ants to a picnic, so extra bandwidth always comes in handy.

Switches and Network Bandwidth

Switches provide you with multiple Ethernet channels, and make it possible to upgrade those channels to higher speeds of operation. Each port on a switch can operate at different speeds, as needed, and is capable of delivering the full bandwidth of the chan-

nel. Examples of switch configurations are provided in Chapter 19. You can link stacks of switches together, for example, to create larger Ethernet systems.

Growth of Network Bandwidth

Today, all computers in the workplace are connected to networks, and everyone in the workplace requires a computer and a network connection. Huge numbers of network applications are in use, ranging from applications that send short text messages to high-definition video. This proliferation of network is leading to a constantly increasing appetite for more bandwidth.

The incredibly rapid growth of the Internet has had a major impact on the traffic flow through local networks. In the past, many computing resources were local to a given site. When major resources were local to a workgroup or to a building, network managers could depend on the 80/20 rule of thumb, which stated that 80% of traffic on a given network system would stay local, and 20% would leave the local area for access to remote resources.

With the growth of Internet-based applications, the 80/20 rule has been inverted. As a result of the Internet and the development of corporate intranets, a large amount of traffic is being exchanged with remote resources. This can place major loads on backbone network systems, as traffic that used to stay local is now being sent over the backbone system to reach the intranet servers and Internet resources and cloud-based services.

Changes in Application Requirements

Not only is traffic increasing, but multimedia applications that deliver streaming audio and video to the user are in common use. These applications place serious demands on network response time. For example, excessive delay and jitter can cause problems with real-time multimedia applications, leading to breakups in the audio and to jerky response on video displays.

Multimedia applications are currently undergoing rapid evolution. Fortunately, modern multimedia applications are designed for delivery over the Web. These applications typically expect to encounter network congestion and packet loss on the Internet. Therefore, they use sophisticated data compression and buffering techniques and other approaches to reduce the amount of bandwidth they require, and to continue working in the presence of low response times and high rates of jitter. Because of this design, these types of multimedia applications will very likely perform quite well even on heavily loaded campus Ethernets.

Designing for the Future

About the best advice anyone can give to a network designer is to assume that you will need more bandwidth, and probably sooner than you expect. Network designers should:

Plan for future growth and upgrades
: The computer business in general, and networking in particular, is always undergoing rapid evolution. Assume that you are going to need more bandwidth when you buy equipment, and buy the best that you can afford today. Expect to upgrade your equipment in the future. While no one likes spending money on upgrades, it is a necessity when technology is evolving rapidly.

Buy equipment with an eye to the future
: Hardware evolution has become quite rapid, and hardware life cycles are becoming shorter. Beware of products that are at the end of their product life cycle. Try to buy products that are modular and expandable. Investigate a vendor's track record when it comes to upgrades and replacing equipment. Look for "investment protection" plans that provide a trade-in discount when upgrading.

Be proactive
: Keep an eye on your network utilization, and regularly store data samples to provide the information you need for trend analysis and planning. Upgrade your network equipment before the network reaches saturation. A business plan, complete with utilization graphs showing the upward trend in traffic, will go a long way toward convincing management at your site of the need for new equipment.

CHAPTER 21
Network Troubleshooting

The best kind of troubleshooting you can do is no troubleshooting at all, and the best way to minimize troubleshooting is to insist on reliable network designs based on conservative practices. On the other hand, there are a lot of components and devices in a network system, and something is bound to eventually go wrong, even in the best of networks.

For those times when your network develops a problem requiring troubleshooting, you need to know how to go about the task of tracking down the failure. There are many ways for things to go wrong in a complex network system. However, the basic approaches to troubleshooting described in this chapter can help you find any problem, no matter how complex the network system may be.

Reliable network design is the best way to avoid network downtime in the first place, so we'll begin this chapter with some guidelines for building a reliable network. We'll also describe two important pieces of information you will need when troubleshooting: network documentation, and baselines of network activity so that you have some idea of normal traffic behavior on your network.

Knowing how to organize the troubleshooting task can help speed the process. Therefore, we will look at the troubleshooting model, including fault detection and fault isolation. These concepts make it possible for you to isolate a problem in any network, big or small.

After looking at the basic troubleshooting concepts, we'll take a tour of the common problems that can be found in the two most widely used cabling systems: twisted-pair and fiber optic. The information in this chapter is based on years of experience in the field, and on real-world reports from network managers at sites all over the globe. Finally, we look at network operation above the level of cables, and describe troubleshooting based on Ethernet frames and high-level network protocols.

Reliable Network Design

One of the best ways to avoid unnecessary network downtime is to make a special effort to design for reliability. Probably the single most important way to improve reliability is to make sure that your network cabling and signaling system meets all standards, and is correctly built using quality components.

Over the years, a number of surveys have found that roughly 70% to 80% of all network failures are related to the network medium. The network medium includes the cables, connectors, and hardware components that make up the signal-carrying portion of an Ethernet system. Many problems with media systems are due to:

- Improperly installed hardware
- The use of incorrect components
- Network designs that violate the official guidelines
- The result of some combination of the above

Ethernet is a mature technology with years of proven multivendor interoperability. In practice, what this means is that you can buy Ethernet equipment from a wide range of vendors, mix it all together, and expect your system to work well. Ethernet equipment is designed to be reliable, and the network devices you buy from vendors will rarely fail.

However, none of this will help you keep your network running if the media system used to link the equipment together is not built correctly. Therefore, the best way to avoid network problems, long troubleshooting sessions, and network downtime is to make sure that your media system is designed and built to be as reliable as possible.

To create the most reliable network, you should:

Design for reliability from the start
: Installation of cables and other hardware represents a major part of the expense and effort in any network installation. Once things are installed, they tend to stay the way they were originally built. Therefore, you really only have one opportunity to do things right: at the beginning, when the network is first being designed and installed.

  Network reliability is a goal that should always be kept in mind when designing and building a network. Reliability is the result of choosing the network topology that provides the most manageable network system, given your resources.

Resist stretching the rules
: Reliability is also improved by choosing quality network components and installing them carefully and correctly, and by resisting the urge to stretch the rules. The specifications contain sufficient engineering margins to allow for some variation between components purchased from different vendors and used together on the

same network. However, a maximum-sized Ethernet system is carefully engineered right out to the last nanosecond of signal delay and jitter budget. Nonstandard equipment, overlong cables, and other such kludges can and will cause problems.

Keep your network designs within the official guidelines, and you will have many fewer problems with your network as time passes and the system grows and expands. To help you accomplish this task, the official guidelines for the various media systems are described in Part II of this book.

Design for future growth
: It's a truism that networks never shrink—they only grow. It can be quite surprising how fast they grow, too. That's why the prudent network designer always tries to accommodate network growth in every network design. You should do this even if present-day users are only thinking about today's needs, and haven't yet thought about how many stations they'll need to support tomorrow.

Avoid "temporary" networks that can become a permanent embarrassment
: It is sometimes tempting to build a temporary lash-up just to get something going until resources can be found to build a "real" network. While this may sound reasonable, it can lead to problems. For one thing, temporary networks have a habit of becoming permanent once users start depending on them to get their work done. Additionally, lash-up networks are usually designed with no thought of future network expansion, making network reliability a real challenge.

Network Documentation

When a network is failing, you want to focus your time on troubleshooting the problem, not documenting the network system. Therefore, one of the most important troubleshooting tools that you can provide for your network is an accurate and up-to-date network map and cable database. Network systems are always growing and changing, so network maps and cable databases require constant updating. Even if you don't always update your documentation, having something on hand is much better than nothing when the network is failing and you have to find out where the problem is located.

There are several drawing packages and database systems sold for network and cable documentation tasks. The high-end packages are expensive, because they are typically based on computer-aided design (CAD) software and include a database for handling large numbers of network devices and cable segments. The mid-range drawing packages are designed for smaller networks and are easier to use and less expensive to buy.

Without documentation, you must begin your troubleshooting with the time-consuming task of finding out how the system is laid out, where the equipment is located, and where the cables go. To speed troubleshooting, your cables should be labeled as discussed in Chapter 15 to make it easier to track the cables down using the information in a cabling database. Without any labeling on the cables, and without a cable database

to list the cables and show how they are laid out, you can spend quite a lot of time hunting for cables and tracing their paths.

Equipment Manuals

Another important set of documents are the equipment manuals. It is often said that the first rule of intelligent tinkering is "Save all the parts." For intelligent networking, we can restate this as "Save all the manuals." You should set up a storage place for manuals, and put every manual you receive into it. Even the single-sheet instructions that sometimes come with small devices like transceivers should be saved.

Having a complete collection of manuals can save you a lot of time when it comes to verifying the correct configuration of a device. It can also save you time when it comes to figuring out what the lights on a given device may mean. When troubleshooting, you often need to know the exact meaning of the troubleshooting lights on equipment; this can sometimes be hard to tell without a manual as the labels used for some lights may be very cryptic. Further, some vendors use their troubleshooting lights to indicate multiple things, depending on the color of the light, or whether the light is constantly lit or flashing.

You should always remember that troubleshooting lights that indicate transmit/receive activity are artificially stretched in length to make them visible to the human eye. The amount of time the lights are lit is quite large compared to the speed of events on the network. For example, a single 64-byte frame will take 51.2 microseconds (μs) to transmit on a 10 Mb/s Ethernet system. This event is typically stretched to about 50 milliseconds (ms) to make it visible to the eye, which makes the length of time the light remains lit approximately 1,000 times longer than the duration of the actual frame transmission.

Therefore, you can only use these lights as a very rough measure of activity. If the network is busy, these lights may be continually lit, which might appear to be an indication of overload or excessive collisions. However, there is no way to accurately determine such problems from these lights, simply because the lights are designed to be on for an artificially long time to make them visible.

System Monitoring and Baselines

When you are trying to find a problem on a network, it can be very useful to know what the normal traffic patterns and error rates look like on that network. By equipping your network with managed switches and SNMP probes, and by regularly polling your equipment to determine traffic levels and error levels, you can create a set of reports that can be stored for future use. When troubleshooting a problem, these reports can be consulted to determine what the normal error rates and traffic rates look like.

Network monitoring packages are available that provide regular polling and report creation for networks. Some of these packages can automatically generate reports and provide them over the Web. This makes it very easy to access the information when you need to find out what the traffic and error profiles have been for a given Ethernet system. A few examples of network monitoring packages are provided in Appendix A.

The Troubleshooting Model

When troubleshooting a network, it helps to have a plan of attack. An effective way to go about troubleshooting a network uses a combination of the scientific method and the technique of divide and conquer. The scientific method of troubleshooting (shown in Figure 21-1) is based on forming hypotheses and testing them. Using your knowledge of the symptoms and of how networks operate, you form one or more hypotheses to explain the behavior you are seeing, and then you perform tests to see if those hypotheses hold up.

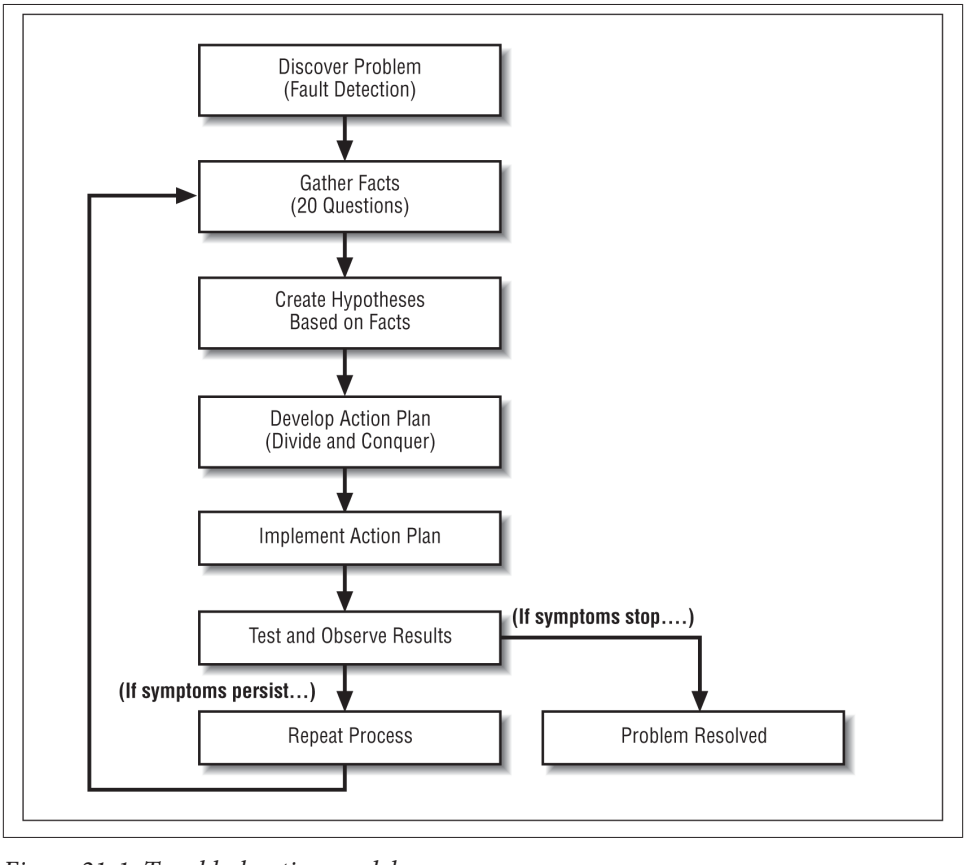


Figure 21-1. Troubleshooting model

The steps are:

Discover the problem
: This is the fault detection stage, in which you are notified of a problem. Notification may be done by automatic fault detection software, or users may call you with a problem report.

Gather facts
: This is a process of acquiring information about the problem. This is rather like the game of "Twenty Questions," in which you ask leading questions to gather information.

Create hypotheses
: Based on the facts you have gathered and your knowledge of how the network functions, you should be able to create one or more hypotheses about the source of the problem. When doing this, you want to make sure that you do not overlook the obvious. Indeed, you want to test the obvious hypothesis first, before spending time on more complex theories. Try to avoid jumping to conclusions, and do not make unnecessary assumptions about the cause of the problem. Make sure that the hypotheses you create can adequately account for the symptoms and other information you have collected.

Develop an action plan
: At this stage, you may have enough information to begin tests of a given device in the network. A test might be as simple as replacing the device with a spare, and then checking to see if the problem is resolved. On the other hand, at this stage you may need to further isolate the problem, in which case your action plan may include some variety of "divide and conquer," or binary search, which is described later in this chapter.

Implement your action plan
: When troubleshooting a problem, try to make only one change at a time. The goal is to eliminate suspected problems one at a time, to limit the number of things you are trying to test and evaluate at any given moment. This way, you can avoid losing track of the problem by trying to evaluate too many things at once.

Test and observe results
: After making a change in the system, you need to test and observe the results, to make sure that you have resolved the problem. If your action plan was based on a binary search, then the test should show you whether the problem is still active. If not, then it is now located in the portion of the network that you have isolated as part of the binary search.

Repeat the troubleshooting process
: If the symptoms still persist, you need to repeat the process until you have resolved the problem.

We'll look at the first stages, detecting and isolating the problem next.

Fault Detection

Fault detection and isolation are at the core of the troubleshooting process. Once you have detected a problem and tracked it down, then you can begin to resolve it. However, the task of fault detection and isolation can be quite complex. The complexity arises from the number of components in the network, the number of users, and the wide range of network applications in use.

Fault detection can cover a range of activities. There are some fault detection systems that will send periodic probe packets to your network equipment to test the reliability of your system. If the equipment doesn't respond after a certain number of tries, it is marked as being down and you are notified. This sort of fault detection is often done with large network management applications based on SNMP. It can also be done more informally using locally developed programs and scripts, or open source software.

For an IP-based network, a fault detection system may use an application called *ping* to send echo request packets to network devices.

The name of the *ping* program comes from the "ping" sound a sonar locator makes when sending out sound waves and detecting objects by their reflection. The *ping* process is much like the sonar system, in that it sends an echo request packet out to a device on the network and receives a "reflection" (echo reply) in response.

When an IP device receives an echo request packet, it will send an echo reply, thereby providing a basic reachability test. Many devices equipped with SNMP management software are also equipped with IP networking software and will respond to echo request packets sent by *ping*. A *ping*-based fault detection system can be designed to send a set of *ping* packets to each IP-equipped device on the network. By sending a series of *ping* packets and keeping track of how many are received, the fault detection system can monitor your network and provide notification when a device fails to respond.

The fault detection stage is often performed by a user who calls up to report that he is unhappy with the way an application is working. These calls can be more difficult to sort out, because there are a wide range of elements involved. These include whether the user's computer has a properly functioning network connection, whether or not the application in use is configured or is being used correctly, and whether the user's computer is functioning properly.

The process of determining if there is a fault, and if so, where the fault is located, is like playing Twenty Questions. This is a game in which your opponent thinks of something,

and you are allowed to ask 20 questions to discover what it is. In the actual game, your opponent must answer truthfully and supplies simple "yes" and "no" answers.

In the network troubleshooting version of Twenty Questions, you are trying to find out what kind of network problem is being reported, and where the problem may be located. Unlike in the game, the person reporting the problem is usually not trying to hide the information from you, and is not limited to *yes* or *no* answers. Still, for many users, the network is a mysterious entity full of unknown devices, and their answers to your questions may not make very much sense. The challenge for the network analyst is to come up with a series of questions that can elicit the information required to define the problem and locate the failing component.

Gathering Information

Symptoms and complaints can help indicate what the problem is and where it is located. You need to collect as much information as you can, by asking the following questions:

- Exactly who is experiencing the problem? Which set of machines are involved, and which network segments?
- Does the problem occur at a certain time?
- How frequently does the problem occur?
- When did the problem first appear?
- When was the exact time of the last occurrence of the problem?
- Has anyone made a change or addition to the network system recently? If so, what was the change and when did it occur?
- Is an error message being generated, and if so, exactly what does it say?
- Is it possible to provide an example of a specific computer that is having the problem, and a detailed description of what the computer is doing when the problem occurs? What application is being run? Is there a specific service involved, and if so which one?
- Can the problem be reliably reproduced?

Fault Isolation

The next stage is to isolate the failure to some portion of the network. When you are involved in isolating a fault on a network, there are several basic approaches you can take. These include determining the network path, duplicating the symptom, and using a binary search to home in on the problem.

Determining the Network Path

There are many components in a network system, and the next task is to figure out which components are in the path of the fault that you are trying to locate. This is where a complete set of network maps and a cable database can save you a great deal of time.

If you are on the trail of a problem in connectivity between a set of computers, or between a client and a server, then you need to know what the network path between those elements looks like. This includes the cable segments, and any switches or routers that might be in the network path. Consider a set of computers connected to a given switch: one that cannot communicate with the rest of the world, while the three other computers connected to that same switch are able to talk to one another. In this case, you need to find out exactly what the network connection for the failing computer consists of. In doing so, you may discover that there is a failing connector on that segment or that someone has just been in the wiring closet that serves the connections on that floor of the building, and may have bumped into the cabling or even moved the cable to another port by mistake.

At one level, a problem such as the one we've just described can look mysterious. Echo packets can be sent to and received from three workstations on that switch with no problems, while another station on the same switch does not respond. Meanwhile, if you are at the keyboard of the "failing" computer, you will find that it is operational. All of this can seem quite odd at first, because all of the computers involved are on the "same" network.

Duplicating the Symptom

If the symptom can be easily duplicated, as in the continuous lack of connectivity for the computer just mentioned, then troubleshooting the problem can proceed rapidly. However, if the problem is intermittent, it can be much more difficult to figure out what may be going wrong. Solid failures at least give you something to work with, so that you can alter various conditions in the network and then retest the system to see if the failure has gone away.

With an intermittent failure, you are faced with a much more difficult troubleshooting task. In this case, you may need to bring some more sophisticated tools to bear. For example, you may wish to install a network probe to provide an independent device for you to monitor.

Some problems seem to occur only when the network is heavily loaded. In that case, you may want to use a network analyzer to inject an artificially high load on the network to see if you can replicate the reported failure. This is best done during off-hours, when people are not trying to use the network to get their work done.

Binary Search Isolation

A major troubleshooting tool is the "divide and conquer" approach, which is more formally known as *binary search*. A binary search isolates the problem through a process of repeatedly cutting the problem area in half, then testing to see if the remaining network is functioning normally.

You start the binary search by disconnecting, or otherwise isolating, one half of the network and then testing the remaining half to see if the problem still exists. This locates the problem on one particular half of the network. You then split the malfunctioning portion of the network in half again, and so on in an effort to reduce the number of possible network sections that could contain the fault.

Depending on how it is done, a binary search can be disruptive, which may limit its usefulness. If a network is solidly down, you may have very little choice but to perform a binary search to isolate the failing portion of the network. For networks that are still functioning but exhibiting some problem, you may want to come in during off-hours to perform a binary search, reducing the impact on the users.

The goal of a binary search is to quickly isolate the failing portion of the network so that you can do more intensive tests to find the failing component. The value of a binary search is that it is the quickest way to find a failing component in a large system. Mathematics tells us that a binary search can find a single component in a system of 1,048,576 components using only 20 tests. However, in the real world, things aren't this simple. For one thing, the speed with which you can perform a binary search depends on how much information you have about the network. Again, a well-documented network is much easier to troubleshoot than one that has little or no documentation at all.

For the best results with a binary search, you need to be looking for a solid failure. Each time you split the network system in half, you need to test the remaining half to see whether the problem still exists. If the problem is intermittent, it can be very difficult to tell if you are making any progress in the binary search. Even with a solid failure, you cannot resolve the problem if you lose track of where it is. Each time you divide the network in half, you need to thoroughly test each half to determine if the problem still exists there or if it is gone and presumably is on the isolated portion of the network.

Dividing network systems

How you divide the network depends on the media system in question and the network components involved. Twisted-pair and fiber optic media systems based on switches can be easy to divide into sections. A binary search can often be performed simply by reconfiguring the switch ports. To do this, you need to investigate the layout of the equipment before conducting a binary search. Then, by shutting off certain switch ports, disconnecting certain cables, or changing the connection points of cables in the system, you can perform a binary search to isolate a problem.

Troubleshooting Twisted-Pair Systems

This section will describe tools and techniques used to troubleshoot twisted-pair cabling systems, including a quick tour of the common problems found in those systems.

Twisted-Pair Troubleshooting Tools

The most commonly used tool for troubleshooting a twisted-pair cabling system is the handheld cable tester. Also called *cable meters* or *pair scanners*, these portable devices come in a wide range of capabilities and prices. A good-quality cable tester can provide a great deal of information about your cabling system and can save you a great deal of time when troubleshooting a cable problem.

Cable testers can be very low-end tools that may only be able to check for the presence of wires on the correct pins of an RJ45-style connector. While these testers can be handy for a quick "go, no go" check of a cable segment, they often don't provide enough information for thorough troubleshooting.

Cable testers can also be much more sophisticated tools that can provide all the basic wiring checks and analyze the signal-carrying characteristics of the wire pairs in your cabling system. High-end cable testers can also test for electrical noise impulses on the cable, as well as testing the cable for length, to make sure that a twisted-pair segment does not exceed the 100 m length recommendation. A high-end cable tester may also come with software that enables the tester to download cable test reports to a computer, allowing you to keep an entire database of cable tests and to print reports when needed.

A high-end cable tester can ensure that a cable segment actually meets the Category 5e or 6A signal specifications from one end of the segment to the other. If you are using 1000BASE-T twisted-pair Gigabit Ethernet over Category 5e cabling, for example, then you need to ensure that your tester complies with the latest testing standards. The testing standards and the Category 5e and 6A cable standards are described in Chapter 15.

Testing for compliance with the Category 5e or 6A signal specifications is a complex task, requiring a number of sophisticated tests to be performed at various signal frequencies. Tests must be made for total signal attenuation at various frequencies, the amount of near-end signal crosstalk (NEXT) that occurs at various frequencies, and so on. Accurate testing requires a high-quality tester. While low-cost testers may provide a series of basic attenuation and crosstalk tests, for example, they may not be able to accurately determine whether a twisted-pair segment meets all of the signaling specifications.

Common Twisted-Pair Problems

This list of common problems is based on years of networking experience, and includes information on network failures reported by network managers at many sites.

Twisted-pair patch cables

Quite a number of problems on a twisted-pair segment can be traced to patch cables. People sometimes attempt to lower costs by making their own patch cables, leading to a number of different failure modes. The most reliable patch cables are those that have been built by a reputable manufacturer under controlled conditions using high-quality materials. Even when purchasing ready-made cables, you should beware of very low-cost patch cables, because they may not meet the cable specifications, or they may be made from inexpensive materials that can result in a loss of signal quality. Problems to look for with patch cables include:

Incorrect wire type
: Patch cords should be built using stranded conductor wire. Using solid conductor wire in a patch cable is a mistake, because it cannot withstand bending or twisting. Eventually, a solid conductor will crack, usually right at the RJ45-style connector on the end of the cable. This leads to intermittent connections, or to open connections on one or more wires that can cause several kinds of problems, including increased bit error rates (cyclic redundancy check, or CRC, errors) and slow network performance. Worse yet, many RJ45 plugs were designed for stranded conductor wire. If used incorrectly on solid conductor cable, these plugs will cut through the solid conductor, virtually guaranteeing an intermittent connection.

  Patch cables for 1000BASE-T Gigabit Ethernet and faster media systems require high-quality stranded cable and high-quality connectors for best performance. The Gigabit and faster systems send signals on all four pairs at high rates, and this intense signaling requires high-quality cable to avoid signal errors. Make sure to use the best-quality patch cables you can find for these links.

Incorrect cable type
: A common mistake in the older 10BASE-T systems was to use a telephone-grade patch, or "silver satin," cable to make the connection between an RJ45 socket in the wall and an Ethernet interface in a computer. Although these cables have stranded wires in them, the stranded wires are very thin and are not twisted together, leading to high signal loss and large amounts of signal crosstalk.

  While a silver satin patch cable may seem to work in a 10BASE-T system, the twisted-pair segment will be experiencing signal errors, causing the network applications to retransmit lost packets, which eventually leads to complaints about slow network performance. Silver satin cables should never be used to make a connection to a standard 10BASE-T transceiver. Silver satin cables will fail to work at all for Fast Ethernet systems.

Incorrectly wired patch cable
: Homemade cables may be incorrectly wired, which is something that can be tested with a cable tester. A cable tester can provide a "wire map" of a cable, showing which

wires are connected to which pins. This can detect whether the cable has a split-pair problem, in which the correct pairing of the wires in the cable is not maintained.

Another problem is homemade patch cords made using a four-pair cable, but with only two pairs (four wires) crimped into place on the connectors at each end. This is sometimes done because older Ethernet media types only need connections on pins 1, 2, 3, and 6. The other four wires are then cut off flush with the end of the cable before the RJ45-style connector is installed. This sort of cable can cause high bit error rates on Fast Ethernet segments. Even if you are not using the other four wires in a cable, all four pairs (eight wires) should be properly terminated in the RJ45-style connector for the best results.

50-pin connectors and hydra cables

While no vendor uses 50-pin connectors on Ethernet switches anymore, a number of old 10BASE-T repeater hubs and switches came with 50-pin connectors to conserve space on the device. This required the use of 25-pair cables containing 50 wires to make a connection between the switch and the wire termination equipment in a wiring closet. If you encounter equipment this old, your best bet is to replace it.

However, if you need to troubleshoot it, then things to look for with 50-pin connectors include:

Loose 50-pin connectors
: The 50-pin connector is not standardized, and connection methods vary. Some vendors used locking clips or Velcro straps, some equipped their connectors with screws, and some had a combination of clips and screws. If not solidly locked down, the 50-pin connector can come loose, often at one end, while still remaining in place. A quick glance might indicate the connector is in place, but might not reveal that the connector is slightly loose at one end, causing an interruption in service or intermittent service to some ports. Always make sure these connectors are firmly in place and locked down.

Multiple disturber crosstalk in hydra cables and 25-pair cables
: For 10BASE-T systems, near-end crosstalk causes the most problems. NEXT occurs when a signal is coupled from a transmitting wire pair into a receive wire pair at the end nearest the transmitter, because that's where the signal is strongest. A typical four-pair twisted-pair cable will support one 10BASE-T connection. However, 25-pair cables and hydra cables were used to support multiple 10BASE-T connections. A hydra cable has a 50-pin connector at one end and is broken out into multiple RJ45 cables and connectors at the other end, leading to the use of the term "hydra," or "many headed," to describe the cable.

  When all of these connections are simultaneously active, it is possible for the multiple signals on the transmit pairs (multiple disturbers) to couple into receive pairs of the cable, causing increased crosstalk and bit errors. This can be difficult to

troubleshoot, because it may only occur when most of the wire pairs are active. Testing this is also difficult because it requires test equipment that can activate all transmit pairs while testing for crosstalk on the receive pairs.

Twisted-pair segment cabling

Twisted-pair patch cables and 50-pin connectors are found on the end of a twisted-pair segment that terminates in a telecommunications closet and connects to patch panels and then to switches. However, most of the twisted-pair cable in a given segment is in the 90 m of cable that travels between the telecommunications closet and the work area in a structured cabling system. This is the cable that gets routed through ceilings and walls, and is terminated at each end with an RJ45-style jack. Things to look for in a twisted-pair segment include:

Excessive amount of untwisting at the wire termination points
: For a segment to meet Category 5e or 6A signal specifications, it must have low levels of signal crosstalk. Crosstalk is reduced when the wires in each pair are tightly twisted together. If the two wires in each wire pair are untwisted too far when the wires are installed in a wire termination point, then excessive crosstalk can result, leading to signal errors on the segment. A high-quality cable tester can determine if a segment meets the crosstalk requirements.

Too many wire terminations
: An excessive number of patch panels or punch-down blocks on a given segment can lead to problems with signal reflections on the wire pairs of that segment. Each connection point in a cable represents some level of impedance mismatch to the flow of signals, and too many connection points can reduce signal strength and cause signal errors. A high-quality cable tester can determine if there is a problem with excessive signal loss or signal reflections.

Incorrect cross-connect wire
: In an old Category 3 system with punch-down blocks used in the segment, a connection may be made between punch-down blocks using a length of telephone-grade cross-connect wire. If this wire is not twisted and rated for Category 3 operation, then the performance of the entire segment will be reduced. Incorrect cross-connect wires should not be a problem in Category 5e or 6A systems because these cabling systems must use only patch panels and patch cords rated for Category 5e and 6A performance in the horizontal link to meet the signal specifications.

Stub cables
: An old Ethernet system based on existing Category 3 or voice-grade telephone wiring could encounter *stub cables* (also known as *bridge taps*), leading to increased signal reflections and noise on the segment. A stub cable is an abandoned telephone cable leading from a punch-down block to some other point in the building. It may have been installed to support an office telephone that later was disconnected. With

telephone systems, stub cables aren't a major problem. However, if the telephone cabling is also used to support 10BASE-T operation, then old stub cables may cause signal reflections and increased bit errors. Again, a high-quality cable tester can indicate whether a segment meets the signal requirements for carrying Ethernet signals.

Troubleshooting Fiber Optic Systems

This section will describe tools and techniques used in troubleshooting fiber optic cabling systems, including a quick tour of the common problems found in this type of system.

Fiber Optic Troubleshooting Tools

A fiber optic media segment works by sending pulses of light over a cable containing glass fibers. It does not use electrical signaling and consequently is immune to electromagnetic interference, greatly reducing the number of things that can go wrong. Furthermore, the tools required for splicing and terminating fiber optic cables are expensive, and are typically used only during the installation of a cabling system. Given this, most sites have their fiber optic systems professionally installed, which also helps ensure that the cable installation will be done properly, minimizing the chances of cabling problems occurring.

One of the simplest and safest tests of a fiber optic link is to connect a fiber optic Ethernet port or outboard fiber optic transceiver to each end of the cable. If the link light comes on, you can assume the segment is working properly. Another simple test is to use an inexpensive fiber optic cable tester based on a light source and light meter to test the segment.

More sophisticated analysis can be done with equipment that sends a calibrated amount of light over the link and measures the exact amount of light loss found from one end of the fiber optic segment to the other. This method can help reveal marginal links where the optical loss is high. For best results, you want to keep the loss as low as possible, so that there is some room for the inevitable small loss of light output and receiver sensitivity that occurs with component aging.

Much more sophisticated analysis can be done with an Optical Time Domain Reflectometer (OTDR). An OTDR is an expensive tool that can measure the amount of light reflected by any discontinuities in the cable. This results in a screen display that can provide a good deal of information to an expert user.

The information includes a measure of total attenuation on the link, as well as pinpointing the exact location of signal loss—an OTDR can determine whether the loss is at a cable splice, a connector, or an excessively tight bend in the cable. A professional

cable installer will typically own an OTDR, which he can use to evaluate the performance of a newly installed system.

Common Fiber Optic Problems

The following is a list of common problems, based on years of networking experience and on information on network failures reported by network managers around the globe:

Connectors incorrectly installed
: Odd as it may sound, it is possible to operate a fiber optic link with connectors that are not firmly in place, as long as enough light can get across the link. There are a variety of fiber optic connectors in use; some use bayonet connections, while others snap into place. No matter which connection method is used, if the connector is not firmly seated, it can cause problems.

  Fiber optic links may work even though a connector is loose or dirty. At some point, however, the connector ends may vibrate far enough apart, or get so much dirt or dust on them that the light levels will be too low and the link will come to a stop. Therefore, it is important to make sure that each fiber optic connector is correctly installed and firmly seated.

Dirty cable ends
: Fiber optic connectors come with dust caps, which must be kept in place until the connectors are used. If the dust caps are left off, the ends of the fiber optic cable inside the connector can accumulate dust and dirt. This will reduce the amount of light that can get through the cable. Finger oils that get onto the ends of the fiber optic cables when they're handled are also a cause for reduced performance in a fiber optic segment. Keep all dust caps in place until a connector is used. Before installing a cable, use a fiber optic cable and connector cleaner to remove dirt and oils from the end of the cable, and from the ends of the fiber optics inside the connectors.

Component aging
: As fiber optic components age, the amount of light a transmitter can send and the optical sensitivity of the receivers is reduced. On very long or marginal links with high loss, this can lead to intermittent failures. One way to troubleshoot this problem is to try new fiber optic transceivers at each end of the link. You can also test the link with a fiber optic power meter to see what the total amount of optical loss may be. This will help determine whether the amount of light carried over the link is marginal.

Data Link Troubleshooting

The next step up from cable troubleshooting is Ethernet frame troubleshooting at the data link layer. Layer 2 of the OSI reference model is the data link layer, which includes the operation of the Ethernet frame. Data link troubleshooting involves monitoring the statistics that switches, interfaces, and management probes can provide about Ethernet frame activity and errors. Frame error reports can be very useful when tracking down a problem because they can help you figure out what kind of problem it might be and where it might be located.

Troubleshooting based on frame statistics has two major components, the first of which is collecting the data. This can be done by using a management station to extract frame statistics from devices on your network. The second component is interpreting the data. It is one thing to collect a bunch of frame statistics, but another to make sense of the statistics you have collected.

The specifics of the task of collecting and interpreting frame statistics vary depending on the number of devices from which you are collecting data. A given troubleshooting session may include looking at the frame statistics on a single switch, in which case the rate of traffic and the presence of any frame errors can provide information on the devices connected to the switch. You can also collect statistics at regular intervals from a large number of switches on your network. In this case you will most likely drown in data if you try to look at all the individual statistics you collect. Some vendors provide network management packages, which can generate network *health reports*. Health reports list only those devices and segments on which a sufficient number of serious errors have been detected. This can save a great deal of time when it comes to interpreting frame statistics retrieved from a large network.

Collecting Data Link Information

Ethernet interfaces located in switches and Ethernet devices like desktop computers can provide statistics and error reports that are useful for troubleshooting. Ethernet management is not required for normal operations, and management capabilities may be optional on these devices. The management information collected from Ethernet devices is described in a set of Internet Request for Comments (RFC) documents. RFCs are available online, and access information for the management RFCs is listed in Appendix A.

The RFCs on Ethernet management include a set of Management Information Base (MIB) documents. A MIB is used to provide a formal description of the information that can be acquired with SNMP. While the MIBs described in the RFCs are formal documents and are not exactly easy to read, they do provide capsule descriptions of each item of management information that can be provided by a given device. A com-

puter equipped with SNMP-based network management software can extract management information from switches and even from Ethernet interfaces in user workstations.

Collecting Information with Probes

A network monitoring device, also called a *probe*, can be installed on a switch port, operating continuously in promiscuous reception mode to provide a set of statistics on the performance of the traffic going through the switch. This feature is an optional management capability on switches.

To use a probe on a switch, you will need packet mirroring capability on the hub. A packet mirror port, also called a SPAN port by one major vendor, is programmed to copy traffic from other ports on the switch to the mirror port for monitoring or analysis.

Network-Layer Troubleshooting

Network-layer troubleshooting refers to troubleshooting at Layer 3 of the OSI reference model, which includes the operations of high-level network protocols. It involves analyzing statistics on high-level network protocol operations and network applications such as the Web and email.

Network-layer analysis can be done using monitoring probes in conjunction with network analysis software. Another network-layer tool is the *protocol analyzer*, which is a device that can capture packets and display the network-layer protocols carried in those packets. A protocol analyzer provides a way to look at network operation above the data link layer.

Network-layer analysis requires knowledge of network-layer protocols and how they operate, as well as high-level applications and how they function. There is a very wide range of applications that use those protocols to send data to one another over Ethernet.

A description of network-layer operations and their analysis is beyond the scope of this book. You should be aware, however, that there are a number of network-layer analyzers on the market. These analyzers can provide information about the operation of the data link layer as well. Depending on the complexity of your network, you may want to invest in a network- layer analysis tool to help troubleshoot problems that may occur.

It is not uncommon to encounter complaints about poor network performance even though the Ethernet data link channel is running fine and is not exhibiting any overloads. Applications such as databases may be functioning slowly because the server is overloaded or improperly configured, or for any number of other reasons that have nothing to do with the operation of the Ethernet.

Some high-level analyzers provide an expert analysis mode that can find and flag problems at the network and application layers, such as excessive protocol retransmissions causing reduced performance.

As we've seen in this chapter, troubleshooting includes a lot of elements and can be complex. The best approach is to avoid as much troubleshooting as you can, by using robust network designs and staying within the rules. But when things fail, having a good grasp of the troubleshooting model and basic troubleshooting techniques can save you a lot of time and effort.

PART VI
Appendixes

The Appendixes provide access to resources for further information, as well as descriptions of older Ethernet technology, including the details of half-duplex mode operation and external transceivers.

APPENDIX A
Resources

The authors maintain a website for Ethernet information that includes a wide range of Ethernet resources. The website includes technical papers on Ethernet and pointers to other web pages with Ethernet information.

The following resources are provided for further information. Resources are listed here as examples only, and no endorsement is implied of any particular company or software package.

Cable and Connector Suppliers

There are many cable and connector suppliers. This list provides a sample of a few major companies whose websites can provide a considerable amount of information about structured cabling and connectors:

Anixter
: A cabling and connector distributor

Belden Cable
: A supplier of coaxial cable, twisted-pair cables, and many other kinds of cables

Corning
: A supplier of fiber optic cabling and components

Hubbell Premise Wiring
: A supplier of structured cabling components

Molex Premise Networks
: A supplier of structured cabling components

Panduit
: A supplier of structured cabling components, including an MPO cable connector, and cable labels

Siemon
: A supplier of structured cabling components

TE Connectivity
: A supplier of cabling and connectors

Cable Testers

Handheld cable testers are used for testing twisted-pair and fiber optic cabling systems. The following listing of vendors of handheld cable testers is provided for information only, and no endorsement of any tester is implied by inclusion in this list:

- EXFO
- Fluke
- JDSU

Cabling Information

The following websites provide information on cable testing issues and cabling systems. The websites for handheld cable testers listed in the previous section are also good sources for cable testing and cabling standards information:

- Cabling-Design.com
- Cabling Installation & Maintenance

Ethernet Jumbo Frames

While the use of jumbo frames has been adopted for specific network designs, such as data centers, there is no official IEEE standard for jumbo frames and therefore no way to guarantee interoperability.

The Internet consists of billions of Ethernet ports operating with the standard maximum frame size of 1,500 bytes. If you want things to work well over the Internet, stick with standard frame sizes. If you're interested in learning more about jumbo frames, however, an IETF draft document on jumbo frame encapsulation is available on the Web.

Ethernet Media Converters

As their name implies, media converters are used to convert from one Ethernet media type to another. These devices can come in handy when you need to connect equipment

that operates at the same Ethernet speed but uses different Ethernet media types. Media converters can also be used to provide an extended-distance link.

A number of vendors provide media converters. The following brief list is necessarily incomplete, and no endorsement is implied by inclusion in the list:

Allied Telesis
: Provides a full line of Ethernet equipment as well as media converters

Canary Communications
: Provides a wide range of media converters

IMC Networks
: Provides media converter products, some of which include SNMP management

Transition Networks
: Provides a range of products to convert copper media to fiber optic

Ethernet OUIs or Vendor Codes

Each manufacturer of Ethernet interfaces acquires an *organizationally unique identifier* (OUI) from the IEEE, which is then used to create the unique 48-bit media access control (MAC) addresses that get assigned to each interface the manufacturer builds (the first 24 bits of the MAC address contain the vendor's OUI). If you know the manufacturer's OUI number (vendor code), you may be able to use that number to identify which computer may be causing network problems. This is not a foolproof mechanism, however, because some vendors may buy their boards from other manufacturers. One networking vendor may also acquire another, at which point it takes over that vendor's OUI.

List of OUIs Maintained by the IEEE

The IEEE maintains a public list of vendor OUIs. The only OUIs listed by the IEEE are ones that vendors give them permission to publish. A vendor may regard the number of OUIs it has requested as competitive information that it would prefer not to have revealed, in which case you will not find that vendor's OUI on the IEEE list. The IEEE also provides instructions for acquiring an OUI.

List of OUIs Compiled by Volunteers

A more complete online list of OUIs has been compiled by the Wireshark network analyzer community, with the help of volunteers from all over the world. This list also contains Ethernet type field identifiers and other information.

Ethernet Bridging and the Spanning Tree Protocol

Useful resources for information on bridging and the Spanning Tree Protocol (STP) include the following:

- Cisco IOS Configuration Guide: "Configuring STP and MST"
  - This chapter of the Configuration Guide describes how to configure the Spanning Tree Protocol (STP) and Multiple Spanning Tree (MST) protocol in Cisco IOS Release 12.2SX.
- Cisco white paper: "Understanding Multiple Spanning Tree Protocol (802.1s)"
  - This white paper documents the differences between Cisco's version of spanning tree, known as Per-VLAN Spanning Tree Plus (PVST+), and the most recent variation of spanning tree developed by the IEEE, known as Multiple Spanning Tree (MST).
- Radia Perlman, *Interconnections: Bridges, Routers, Switches and Internetworking Protocols*, 2nd ed (New York: Addison-Wesley, 1999).
  - An expert's insights into network protocols and how networks function, by the developer of the Spanning Tree Protocol. This book reveals how a protocol designer thinks about networks and protocols.
- Rich Seifert and James Edwards, *The All-New Switch Book: The Complete Guide to LAN Switching Technology* (Hoboken, NJ: Wiley Publishing, Inc., 2008).
  - An exhaustive treatment of LAN switching, from the basics up to the most advanced topics.
- Radia Perlman, "Routing Without Tears, Bridging Without Danger"
  - STP inventor Radia Perlman gave this Google Tech Talk in 2008. In the first half of the presentation, she describes how the Spanning Tree Protocol came about and how basic spanning tree functions. The second half of the talk is about a new bridging protocol that she developed called TRILL.

Layer 2 Network Failure Modes

All network designs have failure modes. For example, Layer 2 networks are vulnerable to traffic loop failures when spanning tree stops working or two ports on a switch with no STP support are connected together.[1] This failure mode and other compatibility issues that have been reported are described in the following resources:

1. This happens more often than you might expect. For some reason, people like to play with Ethernet cables.

- Scott Berinato, "All Systems Down," *CIO*, April 11, 2003.
  - This article in *CIO* magazine documents the failure of a large Layer 2 network design at a major Boston hospital that forced the staff to revert to paper-based operation while the network was repaired. It's rare to find such a useful description of a network failure. Many sites do not publish reports on failures or investigate root causes, and the lack of useful reports on network failure modes makes this report especially valuable.
- John D. Halamka MD, "The CareGroup Network Outage," *Life as a Healthcare CIO*, March 4, 2008.
  - This blog posting provides more details on the Layer 2 network failure at the hospital, and "lessons learned."
- Cisco Systems, Inc., "Troubleshooting Cisco Catalyst Switches to NIC Compatibility Issues," October 2009.
  - This guide from Cisco provides a listing of the major compatibility problems that Cisco customers have encountered with Ethernet interfaces from a variety of vendors.

Cisco Validated Design Guides

Networking vendor Cisco Systems makes available a set of documents covering "validated designs" that include networking designs for a wide range of network environments. While they feature Cisco equipment, these guides contain a lot of useful information on the topics described.

One of the design guides that goes into detail on network design is the "Campus Network for High Availability Design Guide."

Ethernet Switches

There is a large worldwide market for Ethernet switches, with many vendors. Each vendor has a product line aimed at a given market or set of markets.

It would be a major task to provide a comprehensive buyer's guide to the Ethernet switch market. Instead, we will mention just a few of the major manufacturers, which should not be taken as an endorsement of these companies or their equipment.

Manufacturers of switches for the general consumer and small and medium businesses include:

- Dell

- NETGEAR

Vendors of campus, enterprise, data center, and Internet service provider switches include the following:

- Arista
- Cisco Systems
- Hewlett-Packard
- Juniper Networks

Network Protocol Analyzers

There are several network protocol analyzers available. The following references are provided as an example of what a couple of widely used network analyzer products look like and what they include:

Wireshark protocol analyzer
: The Wireshark open source project develops and maintains a sophisticated and very useful network protocol analyzer

Network Instruments Observer
: Network Instruments (now owned by JDSU) provides a complete protocol analysis and performance monitoring system

Network Management Information

The resources listed in this section can be consulted for further information on network management issues:

Multi Router Traffic Grapher (MRTG)
: MRTG is a widely used software package that uses SNMP to monitor network equipment. MRTG data is used to generate a set of graphs that can be viewed with a web browser.

iperf
: *iperf* is an open source application that can stress-test a network with high data rates, as well as provide network performance information. *iperf* runs on multiple platforms.

Simple Network Management Protocol (SNMP)
: Open source SNMP software and information can be found on the Net-SNMP website.

Performance analysis tools
: While somewhat dated, Joseph D. Sloan's *Network Troubleshooting Tools* (O'Reilly, 2001) provides a useful look at a variety of performance analysis issues.

Data communications latency
: This topic is defined in RFC 1242.

Measuring switch latency
: This topic is discussed in RFC 2544.

Latency testing
: The QLogic white paper "Introduction to Ethernet Latency" describes latency testing in detail.

Requests for Comments (RFCs)

When an official standard is developed for the Internet Protocol (IP), it is published in a numbered document called a Request for Comments (RFC).

There are several RFCs that define SNMP MIBs for Ethernet devices. You can find the RFCs that reference Ethernet or switches by searching on the IETF Tools site with the keyword "mib". These two RFCs are provided as examples of what you may find:

- RFC 3635, "Definitions of Managed Objects for the Ethernet-like Interface Types."
- RFC 2613, "Remote Network Monitoring MIB Extensions for Switched Networks Version 1.0." This is the SMON MIB.

Power over Ethernet

Useful resources on Power over Ethernet (PoE) include the following:

- Cisco white paper on PoE and Cisco's vendor-specific extensions
- Cisco guide to troubleshooting PoE contains useful details on PoE design and operation
- The Wikipedia entry on PoE provides information on PoE operation

Standards Documents and Standards Organizations

This book is primarily based on the 802.3 IEEE Ethernet standard. However, there are a number of other standards groups and industry organizations involved in networking.

OSI Model

- The OSI model is an architectural model that describes the set of tasks involved in computer communication as a set of abstraction layers.These layers are also used to organize the set of standards that are developed for computer communication.

BICSI

Building Industry Consultants Service International (BICSI) offers a set of informational publications for cabling professionals.

Fibre Channel Standards

Information on Fibre Channel can be found on the Fibre Channel Industry Association (FCIA) website.

IEEE 802.3 (Ethernet) Standard

The formal IEEE Ethernet standards are a moving target, given that old versions of the standards are continually being updated and new standards are continually being created.

Alternatively, you can purchase a printed version of the latest IEEE standard for Ethernet, published on December 28, 2012.

Information on the 802.3 supplements and working groups can be found on the the IEEE website.

IEEE 802.1 Bridge and Switch Standards

The specifications for basic bridges are found in the 802.1D standard for MAC bridges. The 802.1D standard was extended and enhanced by the subsequent development of the 802.1Q-2011 standard, "Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks."

You can find the published IEEE 802.1 "Bridging and Management" standards online.

For information on active and archived projects for the IEEE 802.1 working group, go to the IEEE website.

Telecommunications Cabling Standards

The Telecommunications Industry Association (TIA) provides a set of widely used structured cabling standards for commercial installations, including TIA-568-C.0, TIA-568-C.1, TIA-568-C.2, TIA-568-C.3, and TIA-568-C.4, as well as a set of specifi-

cations for administering cabling systems (TIA-606-B). The standards are available for purchase on the TIA website.

The International Organization for Standardization (ISO) publishes a cabling standard called ISO/IEC 11801, "Generic cabling for customer premises."

The ANSI/TIA-568 family of cabling standards and the ISO/IEC 11801 cabling standard can also be purchased from Global Engineering.

Other Standards Organizations

Websites for standards organizations and vendor consortiums include:

- American National Standards Institute (ANSI).
- Institute of Electrical and Electronics Engineers (IEEE).
- Internet Engineering Task Force (IETF). The IETF creates engineering standards for the TCP/IP protocol suite.

Switch Performance

Ixia is a vendor of performance analysis tools whose website provides information on how performance is measured. These tools are typically used in large corporate and enterprise networks to monitor and analyze network performance.

Switch Latency

You can find a tutorial on switch latency specifications and how they are measured on the Cisco white paper "Understanding Switch Latency."

Switch and Network Management

There are many network management packages available in the market, and a number of vendors provide some level of network and switch management software for their products. It would be impossible to list all of the network management systems—or even a representative sample—in this short section. The packages listed here are examples of network management packages that are not tied to a single vendor or equipment type:

InterMapper
: This package discovers and documents Layer 2 and Layer 3 networks, and includes NetFlow analysis.

NetBrain
: This network documentation and testing package has the ability to discover and diagram a Layer 2 network.

OpenNMS
: This is a large open source project that provides a management system that continually monitors the state of the network, providing service-level agreement (SLA) reports on network availability and alerting on issues with a sophisticated event and notification management system. This system is designed to scale up to large networks. If you are willing to invest the time and effort to learn how it works, and you have the resources to install and manage a system of this complexity, then OpenNMS can provide an "enterprise-grade" management system.

SolarWinds
: Provides a suite of tools that monitor network performance in switches. These tools use SNMP to provide access to interface counters and other switch information.

Statseeker
: A high-performance traffic monitor that uses SNMP to collect interface counters and switch information every 60 seconds. Statseeker can handle large networks with thousands of interfaces and switch ports.

Traffic Flow Monitoring

Interface counters are useful, but they can't show you much about what the traffic on your network is composed of, or where the traffic is going. NetFlow, IPFIX, and sFlow are systems that provide information on traffic flows; they make it possible to collect information on large flows of traffic and get a look at where your traffic is headed, who is generating the largest amounts of traffic, and so on. Here are some resources that you may find useful if you are considering working with one of these systems:

- The Cisco NetFlow web portal.
- Cisco has also published a white paper on NetFlow.
- For an overview of using sFlow for traffic monitoring, download sFlow's PDF on the subject.

APPENDIX B

Half-Duplex Operation with CSMA/CD

Ethernet began as a local area network technology that provided a half-duplex shared channel for stations connected to coaxial cable segments linked with signal repeaters. In this appendix, we take a detailed look at the half-duplex shared-channel mode of operation, and at the CSMA/CD mechanism that makes it work.

In the original half-duplex mode, the CSMA/CD protocol allows a set of stations to compete for access to a shared Ethernet channel in a fair and equitable manner. The protocol's rules determine the behavior of Ethernet stations, including when they are allowed to transmit a frame onto a shared Ethernet channel, and what to do when a collision occurs.

Today, virtually all devices are connected to Ethernet switch ports over full-duplex media, such as twisted-pair cables. On this type of connection, assuming that both devices can support the full-duplex mode of operation and that Auto-Negotiation (AN) is enabled, the AN protocol will automatically select the highest-performance mode of operation supported by the devices at each end of the link. That will result in full-duplex mode for the vast majority of Ethernet connections with modern interfaces that support full duplex and AN.

However, if you encounter older equipment with interfaces that support only half-duplex, or if the interfaces are manually configured on a 10 or 100 Mb/s switch port to use half-duplex, then you may still encounter a link operating in half-duplex mode.

The 1000BASE-T Gigabit Ethernet system was provided with a half-duplex mode of operation, but given that customers had no need for half-duplex mode, vendors did not bother to implement half-duplex support on Gigabit Ethernet products due to a lack of customer demand.

A misconfigured station or a link where Auto-Negotiation is not working correctly for some reason may also end up operating in half-duplex mode. For that matter, there are no doubt still some functioning Ethernet systems that are based on shared media that operates only in half-duplex mode.

Media Access Control Rules

To begin with, let's look at the rules used for transmitting a frame on a half-duplex shared Ethernet system. When transmitting a frame, the station will encounter the following states:

- When a signal is present on the channel, that condition is called *carrier*.
- When a station attached to an Ethernet wants to transmit a frame, it waits until the channel goes idle, as indicated by an *absence of carrier*.
- When the channel becomes idle, the station waits for a brief period called the *interframe gap* (IFG), and then transmits its frame.
- If two stations happen to transmit at the same time, they detect this "collision" of signals and reschedule their frame transmissions. This occurrence is referred to as *collision detection*.

There are two major things an interface connected to a half-duplex channel must do when it wants to send a frame: it must figure out when it can transmit, and it must be able to detect and respond to a collision. We'll first look at how the interface figures out when to transmit, and then we'll describe the collision detection mechanism.

The rules governing when an interface may transmit a frame are simple:

- If there is no carrier (i.e., the channel is idle), then transmit the frame immediately. If a station wishes to transmit multiple frames, it must wait between successive frames for a period equal to the IFG.

  The IFG is provided to allow a very brief recovery time between frame receptions for the Ethernet interfaces. IFG timing is set to 96 bit times. That is 9.6 microseconds (millionths of a second, μs) for the 10 Mb/s varieties of Ethernet, and 960 nanoseconds (billionths of a second, ns) for the 100 Mb/s varieties of Ethernet.

  If there is a carrier (i.e., the channel is busy), then the station continues to listen until the carrier ceases (i.e., the channel is idle).

  This is known as *deferring* to the passing traffic. As soon as the channel becomes idle, the station may begin the process of transmitting a frame, which includes waiting for the interframe gap interval.
- If a collision is detected during the transmission, the station will continue to transmit 32 bits of data (called the *collision enforcement jam signal*). If the collision is

detected very early in the frame transmission, then the station will continue sending until it has completed the preamble of the frame, after which it will send the 32 bits of jam.

Sending the complete preamble and transmitting a jam sequence guarantees that the signal stays on the media system long enough for all transmitting stations involved in a collision to recognize the collision and react accordingly.

After sending the jam signal, the station waits a period of time chosen with the help of a random number generator and then proceeds to transmit again, starting over at step 1. This process is called *backoff*. The randomly chosen time makes it possible for colliding stations to choose different delay times, so they will not be likely to collide with one another again. Delay times are always in multiples of the worst-case round-trip propagation delay of the network (i.e., the *slot time*).

If the next attempt to transmit the frame results in another collision, then the station goes through the backoff procedure again, but this time the range of backoff times that are used in the random choice process will increase. This reduces the likelihood of another collision and provides an automatic adjustment mechanism for heavy traffic loads.

- Once a 10 Mb/s or 100 Mb/s station has transmitted 512 bits of a frame (not including the preamble) without a collision, then the station is said to have *acquired* the channel. On a properly functioning Ethernet, there should not be a collision after channel acquisition. The 512-bit time value is known as the slot time of the Ethernet channel, and represents the worst-case round-trip propagation delay of the network.

  Once a station acquires the channel and transmits its frame, it also clears its collision counter, which was used to generate the backoff time. If it encounters a collision on the next frame transmission, it will start the backoff calculations anew.

Note that stations transmit their data one frame at a time, and that every station uses this same set of rules to access the shared Ethernet channel for each frame transmitted. This process ensures fair access to the channel for all stations, because all stations must contend equally for the next frame transmission opportunity after every frame transmission. The CSMA/CD MAC protocol ensures that every station on the network gets a fair chance to use the network.

A half-duplex Ethernet operates as a logical signal bus in which all stations share a common signal channel. Any station can use the MAC rules to attempt to transmit whenever it wants to, because there is no central controller. However, for this process to work correctly, each station must be able to accurately monitor the condition of the shared channel. Most importantly, all stations must be able to hear the carrier caused by each frame transmission. In addition, a half-duplex media system must be configured

to allow a station to receive notice of a collision within the carefully specified period of the slot time.

The slot time is based on the *maximum* round-trip signal propagation time of an Ethernet system. The actual round-trip propagation time for a given network system will vary, depending on elements such as the length and type of cabling in use and the number of devices in the signal path. The standard provides specifications for the maximum cable lengths and the maximum number of repeaters that can be used for each media variety. This ensures that the total round-trip time for any given Ethernet built according to the standard will not exceed the maximum round-trip time incorporated into the slot time.

Essential Media System Timing

While signals travel very fast on an Ethernet, they still take a finite amount of time to propagate over the entire media system. The longer the cables used in the media system, the more time it takes for signals to travel from one end of the system to the other. The total round-trip time used in the calculation of the slot time includes the time it takes for frame signals to go through all of the cable segments. It also includes the time it takes to go through all other devices, such as transceiver cables, transceivers, and repeaters.

The maximum lengths for shared-channel half-duplex cable segments are carefully designed so that the essential signal timing of the system is preserved, even if you use maximum-length segments everywhere and build the largest allowable system. The guidelines for each media variety incorporate the essential timing and round-trip signal delay requirements needed to make any half-duplex Ethernet up to the maximum-sized system work properly. The correct signal timing is essential to the operation of the MAC protocol, so let's look at the slot time in more detail.

Ethernet Slot Time

As we've seen, the slot time is used in the CSMA/CD system to set an upper boundary on the timing window during which a collision may occur. By also using the slot time as a basic parameter in the media system design calculations, the standards engineers can guarantee that an Ethernet system will work correctly under all possible legal combinations of standard network components and cable segments.

The total set of round-trip signal delays are summed up in the Ethernet slot time, defined as a combination of two elements:

- The time it takes for a signal to travel from one end of a maximum-sized system to the other end and return. This is called the *physical layer round-trip propagation time*.

- The maximum time required by collision enforcement is the time required to detect a collision and to send the collision enforcement jam sequence. Both elements are calculated in terms of the number of bit times required. Adding the two elements together plus a few extra bits for a fudge factor gives us a slot time that is 512 bit times for 10 and 100 Mb/s systems.

The time it takes to transmit a frame that is 512 bits long is slightly longer than the actual amount of time it takes for the signals to get to one end of a maximum-sized Ethernet and back. This includes the time required to transmit the jam sequence. Therefore, when transmitting the smallest legal frame, a transmitting station will always have enough time to get the news if a collision occurs, even if the colliding station is at the other end of a maximum-sized Ethernet.

The slot time includes the signal propagation time through the maximum set of components that can be used to build a maximum-length network. If media segments longer than those specified in the standard are used, the result will be an increase in the round-trip time, and this could adversely affect the operation of the entire system.

Any components or devices that add too much signal delay to the system can have the same negative effect. Smaller networks, on the other hand, will have smaller round-trip times, so that collision detection will occur faster and collision fragments will be smaller.

Slot Time and Network Diameter

The maximum network cable length and the slot time are tightly coupled. The 512 bit times were chosen as a trade-off between maximum cable distance and a minimum frame size. The total cable length allowed in a network system determines the maximum diameter of that system. The following discusses slot time and network diameter for the three types of Ethernet systems:

Original 10 Mb/s slot time
: In the original 10 Mb/s system, signals could travel through roughly 2,800 meters (9,186 feet) of coaxial cable, transceiver cable, and fiber inter-repeater links and back in 512 bit times, providing a nice, long network diameter.

The much shorter 100 m (328 feet) target length for 10BASE-T twisted-pair segments is based on signal quality limitations, not round-trip timing.

Fast Ethernet slot time
: When the Fast Ethernet standard was developed in 1995, the slot time was kept at 512 bits, because changing the minimum frame length would have required changes in network protocol software, network interface drivers, and Ethernet switches.

However, signals operate 10 times faster in Fast Ethernet, so that a Fast Ethernet bit time is one-tenth of a bit time in the 10 Mb/s original Ethernet system. Because each bit time is only one-tenth as long, that means it is "on the wire" for only one-tenth of the time. As a result, 512 bits will travel over approximately one-tenth of the cable distance with a Fast Ethernet system, compared to original Ethernet. This results in a maximum network diameter of roughly 205 meters (672.5 feet) in Fast Ethernet.

This was considered acceptable because by 1995, most sites were using twisted-pair cabling. The twisted-pair structured cabling standards limit segments to a maximum of 100 meters (328 feet), so the smaller maximum diameter of a half-duplex Fast Ethernet system was not a major hardship.

Gigabit Ethernet slot time
: The Gigabit Ethernet system used a new slot time of 512 *bytes* (4,096 bit times), as described later in this appendix. Given that Gigabit Ethernet equipment that supported the half-duplex mode of operation was never sold, this is of academic interest only.

Use of the Slot Time

The slot time is used in several ways:

- The slot time establishes the maximum upper bound for a station to acquire the shared network channel. Once a station has transmitted a frame for 512 bit times, that is long enough for every station on a maximum-sized shared-channel Ethernet to have heard it, and long enough for any news of a collision to have returned from the farthest end of the network back to the transmitting station.

  At that point the station is assured that it has acquired the channel because (assuming that there has not been a collision) all other stations will have sensed carrier after 512 bit times and will defer to the carrier signal. The transmitting station can now expect to transmit the rest of the frame without a collision. The slot time sets the upper bound on 10 Mb/s and Fast Ethernet channel acquisition to 512 bit times.
- The 512 bits of the slot time also serve as the basic unit of time used by the backoff algorithm to generate a waiting period after a collision has occurred. This algorithm is described later in this chapter.
- A valid collision can only occur within the 512-bit slot time, because once all stations on a network have seen the signal on the medium (*carrier*), they will defer to carrier and won't transmit. Because a valid collision can only happen within the first 512 bits of frame transmission, this also establishes an upper bound on the length of the frame fragment that may result from a collision. Based on this, the Ethernet

interface can detect and discard frame fragments generated by collisions, as the fragments are smaller than 512 bits and are too short to be valid frames.

Slot Time and Minimum Frame Length

Setting the minimum frame length at 512 bits (not including the preamble) means that the data field must always carry at least 46 bytes. A frame that is carrying 46 bytes of data will be 512 bits long, and therefore will not be regarded as a collision fragment. The 512 bits include 12 bytes of addresses, plus 2 bytes used in the type/length field, plus 46 bytes of data, plus 4 bytes of frame check sequence (FCS). The preamble is not considered part of the actual frame in these calculations.

The requirement that each frame carry 46 bytes of data does not impose much overhead when you consider how the data field of a typical frame is used. For example, the minimum length for a typical set of IPv4 headers and TCP headers in a TCP/IP packet is 40 bytes, leaving 6 bytes to be provided by the application sending the TCP/IP packet. If the application only sends a single byte of data in the TCP/IP packet, then the data field needs to be "padded out" with 5 bytes of padding data to provide the minimum of 46 bytes. That's not a severe amount of overhead, and most applications will send enough data that no padding data is needed at all.

Collision Detection and Backoff

Collision detection and backoff is an important feature of the half-duplex CSMA/CD protocol. It's also a widely misunderstood and misrepresented feature. Let's clear up a couple of points right away:

- *Collisions are not errors.* Instead, collisions are a normal part of the operation of an Ethernet LAN. They are expected to happen, and collisions are handled quickly and automatically.
- *Collisions do not cause data corruption.* As we've just seen, when a collision occurs on a properly designed and implemented half-duplex Ethernet, it will happen sometime in the first 512 bit times of transmission. Any frame transmission that encounters a collision is automatically resent by the transmitting station. Any frame less than 512 bits long is considered a collision fragment and is automatically and silently discarded by all interfaces.

It's unfortunate that the original Ethernet design used the word "collision" for this aspect of the Ethernet MAC protocol. Despite the name, collisions are not a problem on an Ethernet. Instead, the collision detection and backoff feature is a normal part of the operation of Ethernet, and results in fast and automatic rescheduling of transmissions.

The collisions counted by a typical Ethernet interface are the ones that occur while that interface is trying to transmit a frame. Collision rates can be significantly higher for very heavily loaded networks, such as a network supporting high-speed computers.

In any case, the thing to worry about is the total traffic load on the network. The collision rate is simply a reflection of the normal functioning of an Ethernet, the rate of collisions seen by a given interface is not significant. Chapter 21 provides more information about measuring the performance of an Ethernet channel.

The Ethernet half-duplex MAC mechanism, with its collision detection and backoff system, was designed to make it possible for independent stations to compete for access to the LAN in a fair manner. It also provides a way for stations to automatically adjust their behavior in response to the load on the network.

Collision Propaganda

The Ethernet collision algorithm is one of the least understood and most widely misrepresented parts of the half-duplex mode of operation. You will sometimes hear that the collision detection and resolution mechanism sets severe limits on the throughput of the system. However, that is not correct. Instead, the Ethernet collision detection and backoff mechanism is a normal part of the operation of an Ethernet. It is a fast and low-overhead way to resolve any simultaneous transmissions that occur on a network system that allows multiple access to a shared channel.

The collision backoff algorithm was also designed to allow stations to respond automatically to varying traffic levels, allowing the stations to avoid one another's transmissions. The collision rate on a properly functioning half-duplex Ethernet segment is simply a reflection of how busy the network is, and of how many stations may be trying to access the channel.

Operation of Collision Detection

Although the stations must listen to the network and defer to traffic (*carrier sense*), it is possible for two or more stations to detect an idle channel at the same time and transmit simultaneously. A collision may occur during the initial part of a station's transmission, which is the 512-bit slot time, also called the *collision window*.

The collision window lasts for the amount of time it takes for the signal from a station to propagate to all parts of the shared channel and back. Once the collision window has passed, the station is said to have acquired the channel. No further opportunity for collision should exist because all other stations can be assumed to have noticed the signal (via carrier sense) and to be deferring to its presence.

Collision detection by media systems

The actual method used to detect a collision is medium-dependent. A link segment medium, such as a twisted-pair or fiber optic cable, has independent transmit and receive data paths. A collision is detected in a link segment transceiver by the simultaneous occurrence of activity on both the transmit and receive data paths.

On a coaxial cable medium, the transceivers detect a collision by monitoring the average DC signal level on the coax. When two or more stations are transmitting simultaneously, the average DC voltage on the coax reaches a level that triggers the collision detection circuit in the coax transceiver. A coaxial transceiver continually monitors the average voltage level on the coaxial cable and sends a collision detect signal to the Ethernet interface when the average voltage level indicates that multiple stations are transmitting simultaneously. This process takes slightly longer than collision detection on link segments; the increased time is included in the calculations used for total signal delay on a 10 Mb/s Ethernet.

Late Collisions

Normal collisions occur during the first 512 bits of frame transmission. If a collision occurs after 512 bit times, then it is considered an error and called a *late collision*. A late collision is a serious error because it indicates both a problem with the network system and causes the frame being transmitted to be discarded.

The Ethernet interface will not automatically retransmit a frame lost due to a late collision. This means that the application software must detect the lack of response due to the lost frame and retransmit the information. Waiting for the acknowledgment timers in the application software to time out and resend the information can take a significant amount of time.

Therefore, even a small number of late collisions can result in slow network performance. Any report of late collisions by devices on your network should be taken seriously, and the problem should be resolved as soon as possible.

Common causes of late collisions

The most common cause of late collisions is a mismatch between the duplex configurations at each end of a link segment. Late collisions will occur if a station at one end of a link is configured for half-duplex operation, and a switch port on the other end of the link is configured for full duplex. As full duplex shuts off the CSMA/CD protocol, a full-duplex interface sends data whenever it wants to. If the full-duplex port or station at one end of the link happens to transmit while the half-duplex station at the other end is sending a frame because it believes it has acquired the channel, the half-duplex station will detect a late collision.

Late collisions can also be caused by problems with the media, such as a twisted-pair segment with excessive signal crosstalk. A twisted-pair transceiver detects a collision by seeing simultaneous traffic on the transmit and receive signal wires. Therefore, excessive signal crosstalk between the transmit and receive wire pairs can cause the transceiver to detect "phantom collisions." If the crosstalk lasts long enough to build up to a level that will trigger the collision detection circuit, a late collision can be generated.

The Collision Backoff Algorithm

After a collision is detected, the transmitting stations reschedule their transmissions using the backoff algorithm to minimize the chances of another collision. This algorithm also allows a set of stations on a shared Ethernet channel to automatically modify their behavior in response to the activity level on the network. The more stations you have on a network, and the busier they are, the more collisions there will be. In the event of multiple collisions for a given frame transmission attempt, the backoff algorithm provides a way for a given station to estimate how many other stations are trying to access the network at the same time. This allows the station to adjust its retransmission rate accordingly.

When the transceiver attached to the Ethernet senses a collision on the medium, it sends a collision presence signal back to the station interface. If the collision is sensed very early in the frame transmission, the transmitting station interface does not respond to a collision presence signal until the preamble has been sent completely. At this point, it sends out 32 bits of jam signal and stops transmitting. As a result, the collision signal will stay on the medium long enough for all other transmitting stations to see it.

The stations involved in transmitting frames at the time of the collision must then reschedule their frames for transmission. The transmitting stations do this by generating a period of time to wait before retransmission, based on a random number chosen by each station and used in that station's backoff calculations.

On busy Ethernet channels, another collision may occur when retransmission is attempted. If another collision is encountered, the backoff algorithm provides a mechanism for adjusting the timing of retransmissions to help avoid congestion. The scheduling process is designed to exponentially increase the range of backoff times in response to the number of collisions encountered during a given frame transmission.

The more collisions per frame, the larger the range of backoff times that will be generated for use by the station interface in rescheduling its next transmission attempt. This means that the more collisions that occur for a given frame transmission, the likelier it is that a station will wait longer before retrying. The name for this scheduling process is *truncated binary exponential backoff*. *Binary exponential* refers to the power of two used in the delay time selection process, while *truncated* refers to the limit set on the maximum size of the exponential.

Operation of the Backoff Algorithm

After a collision occurs, the basic delay time for a station to wait before retransmitting is set at a multiple of the 512-bit Ethernet slot time. Note that a bit time is different for each speed of Ethernet, occupying 100 ns on 10 Mb/s Ethernet and 10 ns on 100 Mb/s Fast Ethernet.

The amount of total backoff delay is calculated by multiplying the slot time by a randomly chosen integer. The range of integers that is generated increases in size after each collision that occurs for a given frame transmission attempt. The interface randomly chooses an integer from this range, and the product of this integer and the slot time creates a new backoff time.

The backoff algorithm uses the following formula for determining the integer *r*, which is used to multiply the slot time and generate a backoff delay:

$0 \leq r < 2^k$

where $k = \min(n,10)$.

To translate this into something closer to English:

- *r* is an integer randomly selected from a range of integers. The value of *r* may range from 0 to one less than the value of 2 to the exponent *k*.
- *k* is assigned a value that is equal to either the number of transmission attempts or the number 10, whichever is less.

Let's walk through the algorithm and see what happens. Imagine that an interface tries to send a frame, and a collision occurs. On the first retry of this frame transmission, the value for 2^k is 2 because the number of tries is one (2 to the exponent 1—2^1—gives us a result of 2). The range of possible integers to choose from for the first retry is set by the equation to greater than or equal to 0, but less than 2.

Therefore, on the first retry after a collision, the interface is allowed to randomly choose a value of *r* from the range of integers 0 to 1. The practical effect is that after the first collision on a given frame transmission, the interface may wait zero slot times (i.e., 0 μs) and reschedule the next transmission immediately. Alternatively, the station may wait one slot time, for a backoff time of 51.2 μs on a 10 Mb/s channel.

If the network is busy due to a number of other stations trying to transmit, and the frame retransmission attempt happens to result in another collision, we're now on try number two. Now, the random number *r* is selected from a set of values that range from 0 to one less than the value of 2 to the exponent of 2 (i.e., 4). As a result, this time around the interface may randomly choose the slot time multiplier from the set of numbers 0, 1, 2, 3. The integer that is chosen is multiplied by the slot time to determine the backoff time before retransmitting.

In other words, if the interface sees a collision on the second attempt to transmit the same frame, it will randomly choose to wait zero, one, two, or three times the slot time before retransmitting again. This expansion in the range of integers is the part of the algorithm that provides an automatic adjustment for heavy network traffic loads and the repeated collisions that they cause.

Choosing a Backoff Time

Notice that the interface is *not* required always to pick a larger integer and develop a larger backoff time after each collision on a given frame transmission. Instead, what happens is that the *range* of integers to choose from gets larger, and the larger set of integers that the interface can choose from includes integers that can generate longer backoff times.

For as long as each frame transmission attempt results in a collision, the process continues with the value of k increasing (and hence, the range of the number r exponentially increasing) up to a maximum of 10.

After 10 retries, the value of k stops increasing, representing the "truncation" part of the algorithm. At this point 2^k is equal to 1,024, so that r is being chosen from the range of numbers from 0 through 1,023. If the frame continues to encounter collisions after 16 tries, the interface will give up. At this point, the interface discards the frame, reports a transmission failure for that frame to the high-level software, and proceeds with the next frame, if any.

On a network with a reasonable load, a station will typically acquire the channel and transmit its frame within the first few tries. Once the station has done so, it clears its backoff counter. If a station encounters a collision on its next frame transmission attempt, it will calculate a new backoff time starting with k=1.

An idle channel is instantly available to the first station that wants to send a frame, so the MAC protocol provides very low access time when the offered load on the Ethernet channel is light, which is typically the case. Stations wishing to transmit frames will experience longer wait times as the load on the channel increases, based on the retransmission times generated by the backoff algorithm.

In effect, each station makes an estimate of the number of other stations that are providing a load on the channel. The backoff algorithm provides a way for stations to do this by allowing the stations to base the estimate on the one property of the network traffic that each station can easily monitor: the number of times that a station has tried to send a given frame and has detected a collision. Table B-1 shows the estimates that a station makes and the range of backoff times that may occur on a 10 Mb/s channel.

Table B-1. Maximum backoff times on a 10 Mb/s system

Collision on attempt number	Estimated number of other stations	Range of random numbers	Range of backoff times[a]
1	1	0…1	0…51.2 µs
2	3	0…3	0…153.6 µs
3	7	0…7	0…358.4 µs
4	15	0…15	0…768 µs
5	31	0…31	0…1.59 ms
6	63	0…63	0…3.23 ms
7	127	0…127	0…6.50 ms
8	255	0…255	0…13.1 ms
9	511	0…511	0…26.2 ms
10–16	1023	0…1023	0…52.4 ms
17	Too high	N/A	Discard frame

[a] Backoff times are shown in microseconds (µs) and milliseconds (ms).

As the table shows, a station exponentially increases its estimate of the number of other stations that are transmitting on the network. After the range reaches 1,023, the exponential increase is stopped, or truncated. The truncation provides an upper limit on the backoff time that any station will need to deal with. It also has another effect, because it means that there are 1,024 potential "slots" for a station to transmit. This number leads to the inherent limit of 1,024 stations that can be supported by a half-duplex Ethernet system.

After 16 tries, the station gives up the frame transmission attempt and discards the frame. At this point, the network is considered to be overloaded or broken—there is no point in retrying endlessly.

All of the numbers used in the collision and backoff mechanism were chosen as part of worst-case traffic and population calculations for an Ethernet system. The goal was to determine reasonable time expectations for a single station to wait for network access. On smaller Ethernets with smaller host populations, the stations detect collisions faster. This leads to smaller collision fragments and more rapid resolution of collisions among multiple stations.

Collision Domains

A useful concept to keep in mind while working with shared-channel Ethernet is the notion of the *collision domain*. This term refers to a single half-duplex Ethernet system whose elements (cables, repeaters, station interfaces, and other network hardware) are

all part of the same signal timing system. In a single collision domain, if two or more devices transmit within the propagation delay time, a collision will occur.

A collision domain may encompass several segments, as long as they are linked together with repeaters, as shown in Figure B-1. A *repeater* is a signal-level device that enforces the collision domain on the segments to which it is connected. The repeater only concerns itself with individual Ethernet signals; it does not make any decisions based on the addresses of the frame. Instead, a repeater simply retransmits the signals that make up a frame.

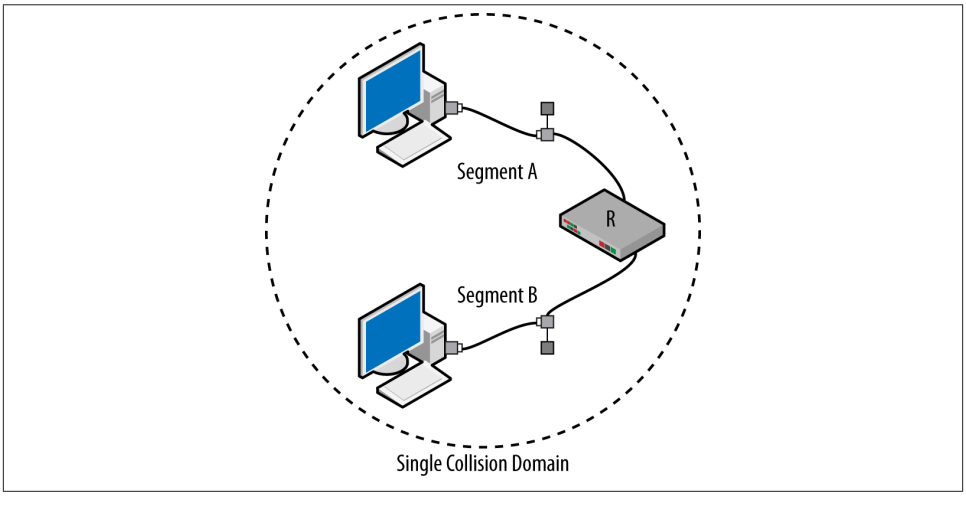


Figure B-1. Ethernet collision domain

Repeaters make sure that the repeated media segments are part of the same collision domain by enforcing any collisions seen on any segment attached to the repeater. For example, a collision on segment A is enforced by the repeater sending a jam sequence onto segment B. As far as the MAC protocol (including the collision detection scheme) is concerned, a repeater makes multiple network cable segments function like a single cable.

On a given Ethernet composed of multiple segments connected with repeaters, all of the stations are involved in the same collision domain. The collision algorithm is limited to 1,024 distinct backoff times. Therefore, the maximum number of stations allowed in the standard for a multisegment LAN linked with repeaters is 1,024. However, that doesn't limit your site to 1,024 stations, because Ethernets can be connected together with packet switching devices, such as switches or routers.

As shown in Figure B-2, the repeaters and computers are connected by means of a *switch*. These Ethernets are in separate collision domains because switches do not forward collision signals from one segment to another.

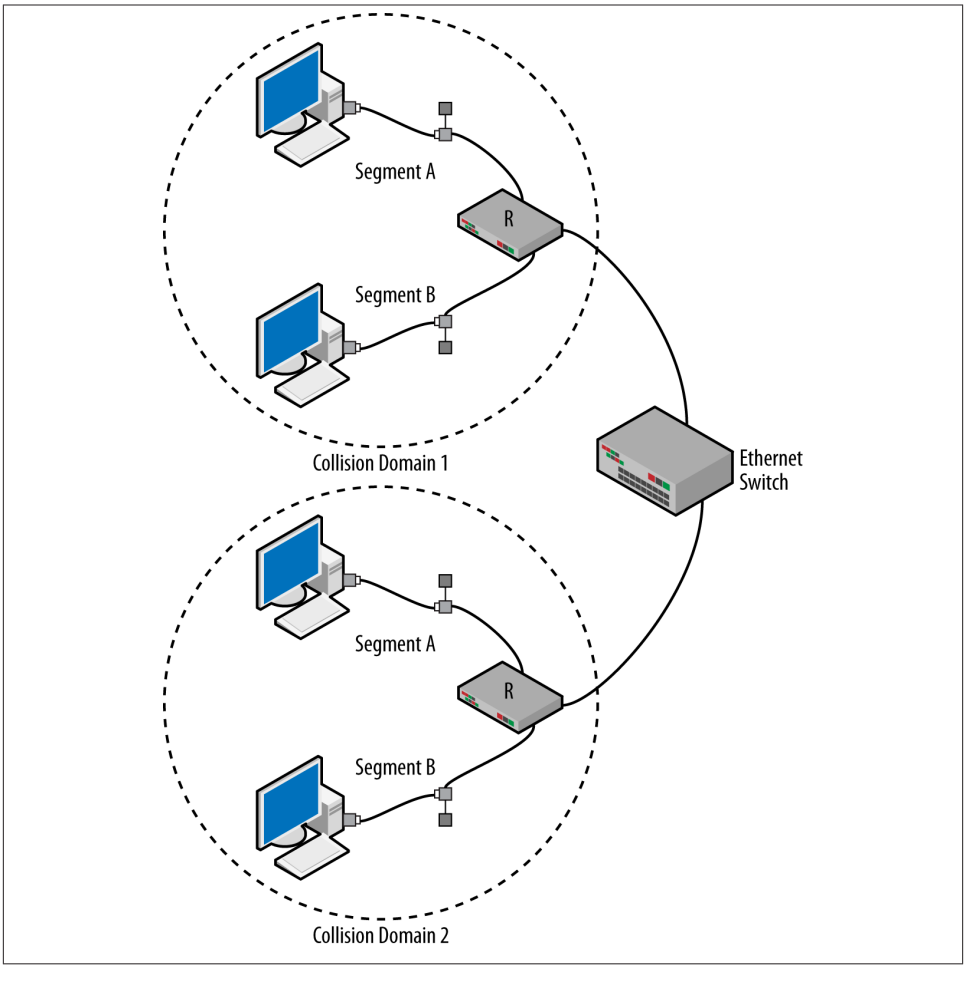


Figure B-2. A switch is used to create separate collision domains

Switches contain multiple Ethernet interfaces, one per port. They operate by receiving a frame on one Ethernet port, moving the data through the switch, and transmitting the data out another Ethernet port in a new frame. When you use a switch to link Ethernets, the linked Ethernets will each have their own separate collision domain. Unlike Ethernet repeaters, switches do not enforce collisions onto their attached segments. Instead, individual Ethernet systems attached to a switch function as separate network systems. Each Ethernet connected to a switch port can be as large as the standard allows. This means that you can use switches to link multiple Ethernets, without being concerned about the sum of all stations attached to all of the Ethernets exceeding 1,024 stations.

As long as switches or routers are used to link the Ethernets, then the round-trip times and collision domains of those Ethernets are kept separate. This allows you to link a whole campus of Ethernet LANs without worrying about exceeding the guidelines for a single collision domain. You can use switches to build large network systems by interconnecting individual Ethernet systems. The operating details of switches are described in Chapter 18.

Ethernet Channel Capture

The Ethernet MAC protocol is a reliable, low-overhead access control system that has proven its worth in millions of Ethernets. However, the MAC protocol is not perfect, and there are aspects of its operation that are not always optimal.

The best-known example of this is an effect called *Ethernet channel capture*. Channel capture results in short-term unfairness, during which a station tends to consistently lose the competition for channel access. There must be one or more stations with a lot of data to send for channel capture to occur, causing them to contend for access to the channel.

The sending station must also be able to continually transmit data at the maximum rate supported by the channel for short periods of time. Continuously sending data in this way sets up the collision and backoff conditions required for channel capture to occur. If the sending station is not capable of continuously sending data at the full rate supported by the channel, then the capture effect will not occur.

Operation of Channel Capture

An example of channel capture works as follows. If you look at all the stations on a channel when several active stations are simultaneously contending for access, you'll expect to see collisions. Each station will possess a nonzero collision counter. As soon as one of the stations acquires the channel and delivers its frame, it clears its collision counter and starts over with a new frame transmission. The rest of the stations trying to transmit will still have nonzero collision counters.

If the *winning* station immediately returns to the channel contention with its backoff parameter in the initial range of 0–1, it has an advantage over the other stations, which have advanced in the algorithm to use wider backoff ranges. The stations with higher collision counters will tend to choose longer backoff times before retrying their frame transmissions, simply by virtue of having a larger pool of values to choose from. The station that wins—which, statistically, is likely to be the one that won the last round, as it has only the low values 0 and 1 to choose from—will return to the fray with a zero collision counter, and therefore tend to continue winning. This station will effectively *capture* the channel for a brief period.

This can only occur if the winning station—which, due to the randomness of the algorithm, could be *any* station—can rapidly and continually transmit data. This requires a high-performance station with a lot of data to send, which was not a common scenario in the early days of Ethernet, when 10 Mb/s was a lot of capacity and the attached stations had relatively low-performance processors. This helps explain why channel capture was first noticed when artificially high network loads were created using performance test software. Another place channel capture may be seen is when a file server is generating large bulk data flows doing backups, while various user machines are trying to access the same channel.

Channel Capture Example

Let's look at a worst-case example of channel capture. Consider the case of two high-performance workstations that both have a significant amount of data to send. Both computers also have high-performance Ethernet interface cards and computer buses that permit them to send frames as fast as the Ethernet channel allows. On their first transmission, they collide and each chooses a backoff of 0 or 1.

Let's say station A chooses backoff 0 and station B chooses 1. In this case, A would transmit its frame while B waits for its one slot time. At the end of A's frame transmission, both A and B are ready to transmit again. They both wait for the interframe gap period, then transmit, collide, and back off. This is A's first collision for this frame transmission attempt, so A will choose a backoff of 0 or 1. Assuming this is B's second collision, then B will sometimes choose a backoff integer between 0 and 3.

If a station (call it station A) sees another station transmitting just before station A's interframe gap timer expires, station A will *still* transmit its frame, and there will be a collision. Seeing another frame on the channel just before the interframe gap expires will *not* prevent a station from transmitting. This ensures fairness. Otherwise, a station with a very slightly faster clock would always "win" the channel because its interframe gap would time out more quickly.

There is a five in eight chance that A will pick a lower number than B, a two in eight chance that they will choose the same number and collide again, and a one in eight chance that B will choose a lower number than A. Therefore, A will most likely win by picking a lower backoff time and transmit its frame. If they pick the same number and collide again, the odds get even worse for B.

Because we're assuming that these two stations have a lot of data to send, the process will then repeat—only this time it is again A's first attempt to transmit a new frame, while poor station B is on attempt number three for its frame transmission. Each time station A succeeds, station B's disadvantage increases. Once A has transmitted three or four frames, it will pretty much be able to transmit at will. B will continue losing the

channel contentions until B's transmission attempts counter reaches 16, at which point the station discards the frame and starts over. At this point A and B are back on even terms, and the contention is fair again.

Long-Term Fairness

If you look at the channel arbitration during the 16 retries that B makes trying to send a frame, this system appears to be unfair. But over a period of several minutes, both A and B will get about equal shares of the channel, but sometimes B will be the winner that gets to send a burst of packets.

For most stations, it does not appear that channel capture often results in a frame discard. Instead, the typical network application will tend to have less than 16 frames' worth of data to send at any given time, and the channel capture will be of shorter duration.

Stations with slower network interfaces that cannot sustain a constant train of back-to-back frame transmissions will tend to break up the capture effect. As soon as a station pauses in its transmission attempts, another station will be able to access the channel. Further, networks with high-performance computers are often segmented into smaller pieces using switches, and the smaller populations on these segmented network channels are less likely to get involved in channel capture.

To really see channel capture at work may take an artificially high load, which is why channel capture may be seen with applications that test network throughput using a constant stream of data being sent between network test programs. For example, if channel capture occurs while measuring network performance using IP software, you could try setting the size of the "window" in the IP network software down to something around 8 KB. This has the effect of reducing the total number of packets that can be sent at any given time, while not making any significant difference in throughput over a 10 Mb/s channel. This, in turn, breaks up the capture effect and allows the network test program to provide the high throughput results that everyone was expecting.

A Fix for Channel Capture?

Channel capture was intensively studied because of the potential bottlenecks it could create. Subsequently, a mechanism called the Binary Logarithmic Arbitration Method (BLAM) was developed as a way to avoid channel capture. The BLAM mechanism eliminates the capture effect by changing the backoff rules to make access to the channel more fair in the short term. The result is to smooth the flow of packets on a busy network.

The BLAM algorithm was designed to be backward compatible with existing Ethernet interfaces, so that mixed networks with standard Ethernet and BLAM-equipped interfaces could interoperate. A complete description of Ethernet channel capture and the

BLAM algorithm can be found in a paper by Mart L. Molle titled "A New Binary Logarithmic Arbitration Method for Ethernet."[1]

An IEEE project called 802.3w was launched to study the implementation of BLAM, and to decide whether to add BLAM to the official Ethernet standard. For a variety of reasons, it was decided not to standardize BLAM. For one thing, vendors were nervous about deploying a new Ethernet operational mode into the field. Despite simulations and lab tests of BLAM, the new algorithm had not been deployed extensively. What if unforeseen complications occurred? There were many millions of Ethernet nodes already in use at the time, which made vendors understandably conservative when it came to changing how Ethernet worked.

Meanwhile, reports from the field indicated that most customers were not encountering the capture effect, so it appeared that vendors would be going to significant effort and risk to solve a relatively rare problem. Many networks were already being segmented with switches, thereby limiting the number of machines contending for access on any given channel and making it harder for channel capture to occur. Further, on the increasingly popular higher-speed Ethernet systems, it was more difficult for channel capture to occur. A given application and set of stations that can cause channel capture on a 10 Mb/s network has to work 10 times as hard to capture the channel on a 100 Mb/s network.

For that matter, the high-performance computers and server systems in newer networks were typically connected directly to an Ethernet switch, with full-duplex operation enabled on the link. Full-duplex mode does not use the Ethernet MAC protocol, so the capture effect can never occur on these links. In the end, the shrinking benefits of adding BLAM appeared to be outweighed by the risks, and the BLAM spec was shelved.

Gigabit Ethernet Half-Duplex Operation

Currently, all Gigabit Ethernet equipment is based on the full-duplex mode of operation described in Chapter 4. No vendors ever provided equipment that supported the half-duplex mode for Gigabit Ethernet operation. Nonetheless, a half-duplex CSMA/CD mode for Gigabit Ethernet was specified, if only to ensure that Gigabit Ethernet met what were the requirements at that time for inclusion in the IEEE 802.3 CSMA/CD standard. For the sake of completeness, a description of Gigabit Ethernet half-duplex mode is provided here.

1. Technical Report CSRI-298, April 1994 (revised July 1994), Computer Systems Research Institute, University of Toronto, Toronto, Canada.

Gigabit Ethernet Half-Duplex Network Diameter

A major challenge for the engineers writing the Gigabit Ethernet standard was to provide a sufficiently large network diameter in half-duplex mode. As we've seen, the maximum network diameter (i.e., cable distance) between any two stations largely determines the slot time, which is an essential part of the CSMA/CD MAC mechanism.

Repeaters, transceivers, and interfaces have circuits that require some number of bit times to operate. The combined set of these devices used on a network takes a significant number of bit times to handle frames, respond to collisions, and so on. It also takes a small amount of time for a signal to travel over a length of fiber optic or metallic cable. All of this is summed into the total timing budget for signal propagation through a system, which determines the maximum cabling diameter allowed when building a half-duplex Ethernet system.

In Gigabit Ethernet, the signaling happens 10 times faster than it does in Fast Ethernet, resulting in a bit time that is one-tenth the size of the bit time in Fast Ethernet. Without any changes in the timing budget, the maximum network diameter of a Gigabit Ethernet system would be about one-tenth of that for Fast Ethernet, or in the neighborhood of 20 meters (65.6 feet).

Twenty meters is usable within a single room, such as a machine room equipped with a set of servers. However, one of the goals for the half-duplex Gigabit Ethernet system was to support a large enough half-duplex network diameter to reach from a Gigabit Ethernet repeater hub to the desktops in a standard office building. Desktop cabling for office buildings is typically based on structured cabling standards, which require the ability to reach 100 meters from a hub port. This means that the total network diameter may reach a maximum of 200 meters when connecting two stations to a Gigabit Ethernet repeater hub.[2]

Looking for Bit Times

To meet the 200 meter half-duplex network diameter goal, the designers of the Gigabit Ethernet system needed to increase the round-trip timing budget to accommodate longer cables. If it were somehow possible to speed up the internal operations of devices such as repeaters, it was thought that it might be possible to increase the bit timing budget. The idea was that you could save some of the bit times that are consumed when signals are sent through repeater hubs. Unfortunately, it was not possible to produce repeaters and other components with one-tenth of the delay of equivalent Fast Ethernet devices.

2. Repeaters can only be used in a half-duplex shared Ethernet channel. By definition, anything connected to a repeater hub must be operating in half-duplex mode.

The next place people thought of looking to save bit times was in the cable propagation delays. However, it turns out that the signal propagation delay through the cables cannot be reduced, as the delay is fundamentally based on the speed of light (which is notoriously difficult to improve).

Another way to gain time and achieve longer cable distance, was in the minimum frame transmission specification. If the minimum frame time were extended, then the Ethernet signal would stay on the channel longer. This would extend the round-trip time of the system and make it possible to achieve the 200 meter goal for the cabling diameter over twisted-pair cable. The problem with this scheme, however, was that changing the minimum frame length would make the frame incompatible with the other varieties of Ethernet, all of which use the same standard minimum frame length.

Carrier Extension

The solution to this conundrum was to extend the amount of time occupied by the signal associated with a minimum frame transmission, without actually modifying the minimum frame length or other frame fields.

As shown in Figure B-3, Gigabit Ethernet does this by extending the amount of time a frame signal is active on a half-duplex system with a mechanism called *carrier extension*. The frame signal, or *carrier*, is extended by appending non-data signals called *extension bits*. Extension bits are used when sending short frames, so that the frame signal stays on the system for a minimum of 512 bytes (4,096 bit times), which is the new slot time for Gigabit Ethernet. This slot time makes it possible to use longer cables, and is also used in the collision backoff calculations on Gigabit Ethernet systems.

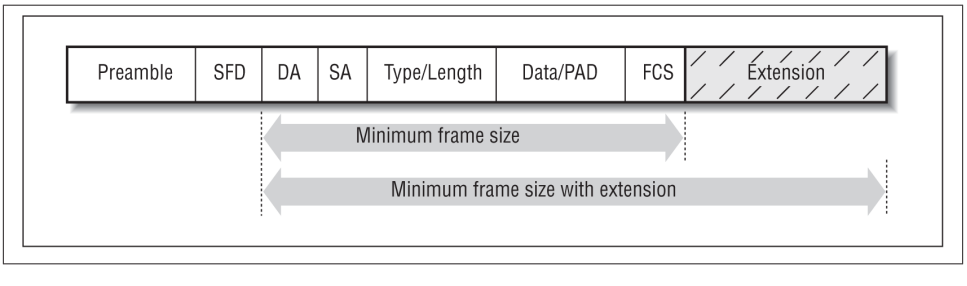


Figure B-3. Carrier extension

The use of carrier extension bits assumes that the underlying physical signaling system is capable of sending and receiving non-data symbols. Signaling for all fiber optic and metallic cable Gigabit Ethernet systems is based on signal encoding schemes that provide non-data symbols that will trigger carrier detection in all station transceivers. This makes it possible to use these non-data symbols as carrier extension bits without having the extension bits confused with real frame data.

With carrier extension, a minimum-size frame of 64 bytes (512 bits) is sent on a Gigabit Ethernet channel along with 448 extension bytes (3,584 bits), resulting in a carrier signal that is 512 bytes in length. Any frame less than 4,096 bits long will be extended as much as necessary to provide carrier for 4,096 bit times (but no more).

Carrier extension is a simple scheme for extending the collision domain diameter; however, it adds considerable overhead when transmitting short frames. A minimum-size frame carrying 46 bytes of data is 64 bytes in length. Carrier extension adds another 448 bytes of non-data carrier extension bits when the frame is transmitted, which significantly reduces the channel efficiency when transmitting short frames.

The total impact on the channel efficiency of a network will depend on the mix of frame sizes seen on the network. As the size of the frame being sent grows, the number of extension bits needed during transmission of the frame is reduced. When the frame being sent is 512 bytes or more in length, no extension bits are used. Therefore, the amount of carrier extension overhead encountered when sending frames will vary depending on the frame size. Carrier extension is only specified for use in half-duplex Gigabit Ethernet. In full-duplex mode, the CSMA/CD MAC protocol is not used, which removes any concern about the slot time. Therefore, full-duplex Gigabit Ethernet links do not need carrier extension and are able to operate at full efficiency.

Frame Bursting

The Gigabit Ethernet standard defines an optional capability called *frame bursting* to improve performance for short frames sent on half-duplex channels. This allows a station to send more than one frame during a given transmission event, improving the efficiency of the system for short frames. The total length of a frame burst is limited to 65,536 bit times plus the final frame transmission, which sets a limit on the maximum burst transmission time.

Figure B-4 shows how frame bursting is organized. The first frame of the transmission is always sent normally, so that the first frame is sent by itself and will be extended if necessary. Because collisions can occur only within the first slot time, only this frame can be affected by a collision, requiring it to be retransmitted. This frame may encounter one or more collisions during the transmission attempt.

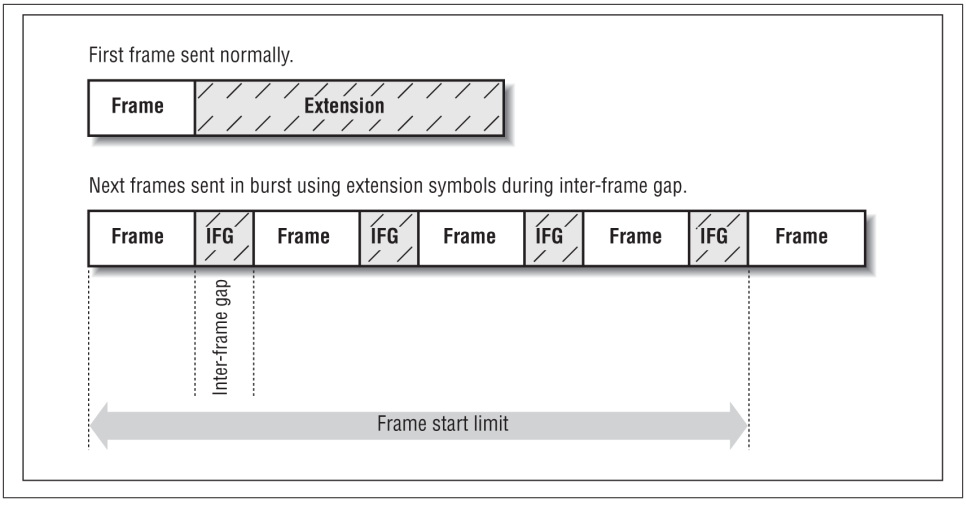


Figure B-4. Frame bursting

However, once the first frame (including any extension bits) is transmitted without a collision, then a station equipped with frame bursting can keep sending additional frames until the 65,536 bit time burst limit is reached. To accomplish this, the transmitting station must keep the channel from becoming idle between frame transmissions. If the station were to become idle during frame transmission, other stations would try to acquire the channel, leading to collisions.

The frame-bursting station keeps the channel active by transmitting special symbols, understood by all stations to be non-data symbols, during the interframe gap times of the frames. This causes all other stations to continue to sense carrier (activity) on the channel. This, in turn, causes the other stations to continue deferring to passing traffic, allowing the frame-bursting station to continue sending frames without concern that a collision will occur.

In essence, the first frame transmission clears the channel for the subsequent burst frames. Once the first frame has been successfully transmitted on a properly designed network, the rest of the frames in a burst are guaranteed not to encounter a collision. Frames sent within a burst do not require extension bits, even if they are shorter than 4,096 bit times. The transmitting station is allowed to continue sending frames in a burst until the *frame burst limit* (FBL) is reached, at which point the last frame in the burst is sent.

For short frames, the optional frame-bursting mechanism can improve the utilization rate of the channel. However, this can only occur if the station software is designed to take advantage of the ability to send bursts of frames. Without frame bursting, the throughput on a half-duplex Gigabit Ethernet channel is less than twice that of a Fast Ethernet channel for minimum-length frames. With frame bursting, the throughput of

a Gigabit Ethernet channel is slightly over nine times that of the Fast Ethernet system for a constant stream of minimum-length frames.

Frame bursting and shared-channel efficiency

Without frame bursting, the channel efficiency is low for a Gigabit Ethernet channel carrying a constant stream of 64-byte (512-bit) frames. It requires an overhead of one slot time to carry the 512 bit minimum-size frame, which is 4,096 bit times in the Gigabit Ethernet system. To this we add 64 bit times of preamble and 96 bit times of interframe gap, for a total of 4,256 bits of overhead. Dividing the 512-bit payload by the overhead of 4,256 bits reveals a 12% channel efficiency.

With frame bursting, the Gigabit Ethernet channel is considerably more efficient for small frames because a whole series of frames can be sent in a burst without overhead for the slot time, once the channel is acquired. Theoretically, you could send 93 small frames in a single burst, with a channel efficiency of over 90%. However, in the real world, it is unlikely that the smallest possible frames will dominate the traffic flow. Nor is it likely that a given station will have so many small frames to send that it can pack a frame burst full of them on a constant basis.

Remember, these limits only occur in half-duplex mode due to the round-trip timing requirements. Carrier extension is not needed in full-duplex mode because full-duplex mode does not use CSMA/CD and is unconcerned about round-trip timing. The ability to send a burst of frames is an inherent characteristic of a full-duplex channel. A full-duplex Gigabit Ethernet system can operate at the full frame rate for all frame sizes, or 10 times faster than a full-duplex Fast Ethernet system. Gigabit Ethernet performs quite well, particularly because the vendors of Gigabit Ethernet equipment support only the full-duplex mode of operation.

APPENDIX C

External Transceivers

In this appendix, we describe two external transceivers and transceiver interfaces that were once widely used but are no longer sold as new equipment. These are the AUI cable and external MAU for 10 Mb/s systems, and the MII cable and external PHY for 100 Mb/s systems. The equipment described here is obsolete and no longer used in new installations. We will use the present tense to describe the operation of these components, but keep in mind that this information is included solely for the sake of completeness.

The attachment unit interface (AUI) cable, also called a transceiver cable, was developed as part of the original 10 Mb/s Ethernet system. New media systems based on twisted-pair and fiber optic link segments were later invented for the 10 Mb/s system, and in the early 1990s the 10BASE-T twisted-pair system became the most widely implemented networking system.

The AUI makes it possible to connect an Ethernet interface to any one of the several 10 Mb/s media systems, while isolating the interface from any details of the specific media system in use. The development of the AUI as a medium-independent attachment was actually a side effect of the design of the original thick coaxial cabling system, which requires the use of external transceivers connected directly to the coax cable.

The need to provide a connection between the Ethernet interface electronics in the station and the external transceiver located on the coaxial cable was what led to the development of the AUI. The development of the AUI, in turn, made it possible to develop other cabling systems for 10 Mb/s Ethernet without requiring any changes in the Ethernet electronics in the station.

The Data Terminal Equipment

A station, more formally called *data terminal equipment* (DTE) in the standard, is a unique addressable device on the network that serves as an originating or terminating

point for data. For example, each Ethernet-equipped computer or port on a switch is a DTE, as each is equipped with an Ethernet interface. The Ethernet interface contains the electronics needed to perform the media access control (MAC) functions required to send and receive Ethernet frames over the Ethernet channel.

Ethernet ports on repeater hubs are *not* DTEs and are not equipped with addressed Ethernet interfaces. A repeater port is connected to an Ethernet media system using standard components, such as a transceiver. However, repeater ports operate at the individual bit level for Ethernet signals, amplifying and retiming signals as they move through the repeater so that they can travel from one segment to another. Repeater ports do not contain Ethernet MAC-level interfaces and do not operate at the level of Ethernet frames.

Figure C-1 illustrates how the AUI is used in a 10 Mb/s system. This figure shows a set of components that can be used to implement a twisted-pair 10 Mb/s Ethernet connection. Both internal and external transceiver connections are shown.

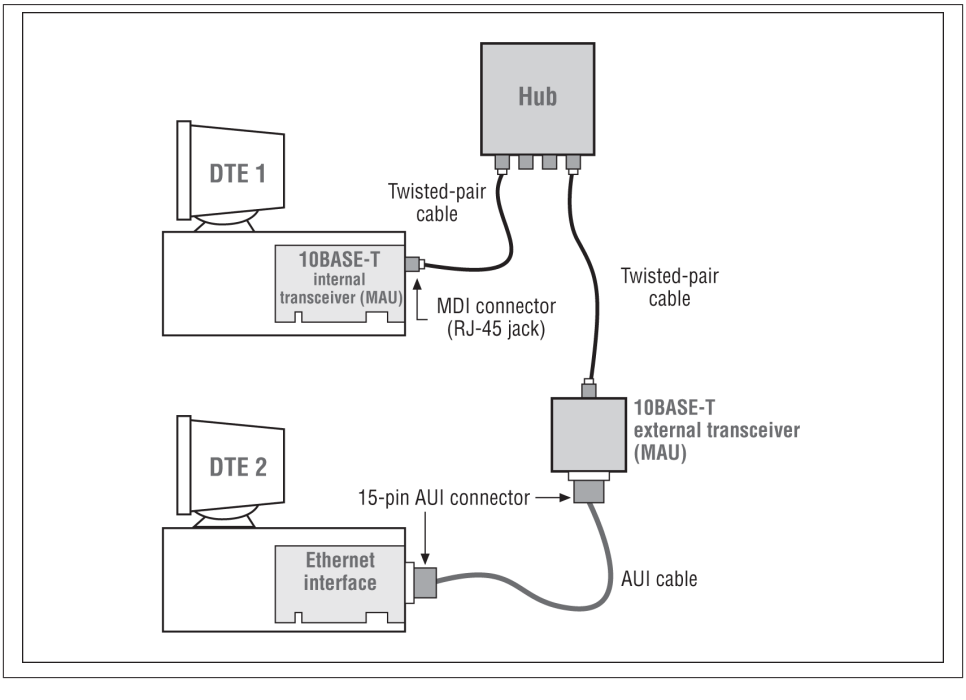


Figure C-1. AUI connection for a 10 Mb/s Ethernet system

The Attachment Unit Interface

The AUI is the medium-independent attachment used in older Ethernet systems to connect to one of several media systems. In Figure C-1, DTE 2 has an Ethernet interface equipped with a 15-pin AUI connector and external transceiver that, in turn, is connected to a twisted-pair cable.

The AUI connector on DTE 2 could be used to connect to several 10 Mb/s Ethernet media systems, by using the appropriate external transceiver. Figure C-1 also shows DTE 1, which has a built-in 10BASE-T transceiver. Because this station does not have a 15-pin AUI connector, it cannot be connected to any other media system; it connects only to a twisted-pair cable.

The 15-pin AUI connector provides an external transceiver connection for a station. This connector provides power to the external transceiver and a path for Ethernet signals to travel between the Ethernet interface and the media system. The AUI connector uses a slide latch mechanism to make an attachment between male and female 15-pin connectors.

The AUI Slide Latch

Rich Seifert, a developer of the IEEE standards and one of the primary design engineers who worked on the design of the original 10 Mb/s Ethernet system, has this to say about the sliding latch connectors:

> *Personally, I would have saved every Ethernet user a lot of grief by not specifying the dreaded slide latch connector (used on the cable between the station and the transceiver). We really had good intentions. I was fed up with the RS-232C connectors that fell off because the tiny screwdriver necessary to tighten them down was never handy. I just didn't realize that the slide latch was so flimsy and unreliable until it was too late. Ethernet installers around the world must curse me every day.*[1]

As you can probably guess from Seifert's remarks, the slide latch connectors used on 15-pin AUI connectors were a source of problems for installers of the early Ethernet systems. In fact, the slide latch was probably the most disliked piece of hardware in the entire 10 Mb/s Ethernet system. That's because it is the part just about every user of AUI-based equipment encountered, and at the same time it was the part that could be voted "least likely to succeed" given the problems caused by poor design and installation.

Figure C-2 shows what a slide latch looks like in the open and closed positions before being installed on the locking posts of the mating 15-pin connector. The view is from the end of the transceiver cable, looking at the connector "end on," and shows the screws that hold the sliding latch clip in place on the end of the connector. The screws are

1. *Rich Seifert, "Ethernet: Ten Years After," Byte 16:1 (January 1991), 319.*

installed in the connector end and do not move. The latch assembly fits underneath the heads of the screws with a small amount of room left to allow for movement, which lets the latch slide back and forth.

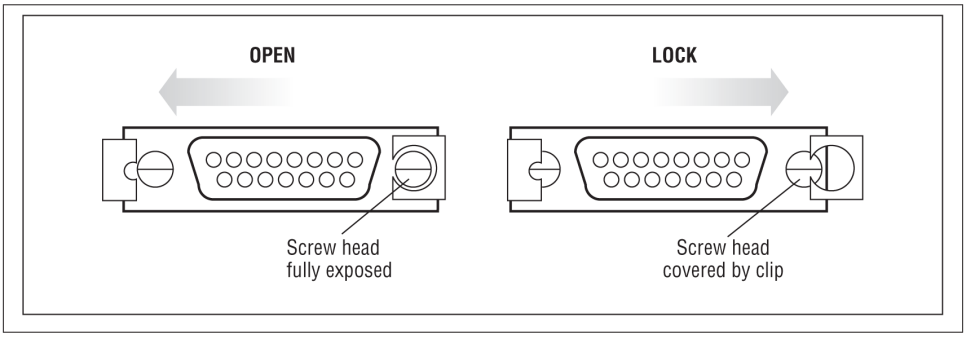


Figure C-2. Sliding latch mechanism

The clips at each end of the latch clip onto the heads of the locking posts provided on the mating 15-pin connector—that's what holds the two connectors together. If everything about the hardware is correct, the latch is fairly hard to move back and forth and "snaps" into the open or closed position. If the hardware is not correct, the latch is loose and will move easily from side to side.

Problems with the sliding latch

Despite the best efforts of the IEEE 802.3 committee, some vendors did not follow the specifications carefully and did not install the slide latch connectors properly on their equipment and cables. To compound the problem, there was wide variability in the quality of the slide latches, with some vendors using lightweight slide latch hardware, resulting in transceiver cables that fell off easily.

Although the slide latch connector may not have been the world's mightiest way to connect things, it could result in a quick and secure attachment when properly installed. A good-quality slide latch connection could be made solid enough to allow you to move the connected computer around by tugging, or more likely tripping over, a transceiver cable. Many vendors managed to do the job right, resulting in very reliable network connections.

On the other hand, when the slide latch connector was bad, it was horrid. Slide latches that were made out of lightweight metal that bent easily could be nearly impossible to get onto the locking posts. If the vendor had also installed the 15-pin connector incorrectly, then the situation was grim. A few vendors even mounted the 15-pin connector just behind the metal frame of the computer, which resulted in the locking pins of the 15-pin connector being located slightly too far away from the surface of the slide latch

when they were connected. That, in turn, led to slide latch connectors that never made a tight fit and that fell off at the slightest provocation.

AUI Signals

The transceiver cable carries a set of signals between the external transceiver and the Ethernet interface. Figure C-3 lists the signals provided by the 15-pin AUI connector. The signals sent by the transceiver and received by the Ethernet interface are low-voltage differential signals. There are two wires for each signal, with one wire for the positive (+) and one wire for the negative (–) portions of the signal. The voltage level on these wires varies from +0.7 volts to –0.7 volts, providing a nominal total of 1.4 volts peak-to-peak for the entire signal.

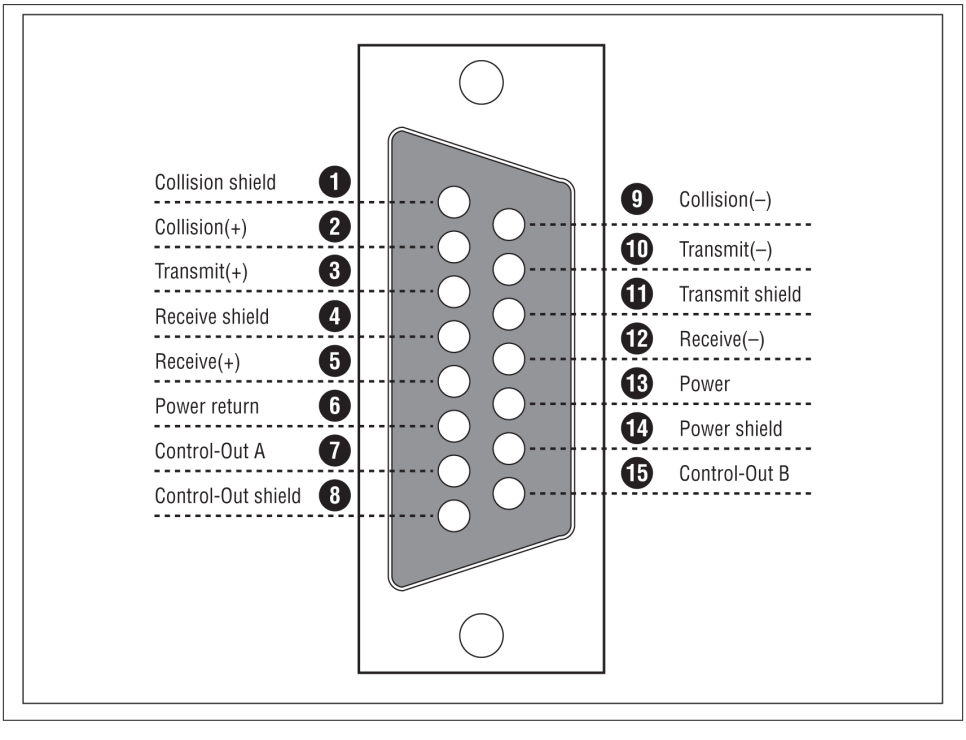


Figure C-3. AUI connector signals

The "Control-Out" signal was provided to send optional control signals from an Ethernet interface to the transceiver. This option was never implemented by any vendor and was not used.

The AUI Transceiver Cable

The 10 Mb/s transceiver cable, formally known as an AUI cable, is built like an electrical extension cord: there's a plug (*male connector*) on one end and a socket (*female connector*) on the other. The transceiver cable carries three data signals between a 10 Mb/s Ethernet interface and the external transceiver:

- *Transmit data*, from the Ethernet interface to the transceiver
- *Receive data*, from the transceiver to the interface
- A *collision presence signal*, from the transceiver to the interface

Each signal is sent over twisted-pair wires. Another pair of wires is used to carry 12-volt DC power from the Ethernet interface to the transceiver. The standard transceiver cable uses fairly heavy-duty stranded wire to provide good flexibility and low resistance.

The AUI transceiver cable, shown in Figure C-4, is equipped with a 15-pin female connector on one end that is provided with a sliding latch; this is the end that is attached to the outboard transceiver. The other end of the transceiver cable has a 15-pin male connector equipped with locking posts; this is the end that attaches to the Ethernet interface. Some 15-pin AUI connectors on Ethernet interfaces have been equipped with screw posts instead of the sliding latch fastener described in the standard, requiring a special transceiver cable with locking screws on one end instead of sliding latch posts.

The AUI transceiver cable described in the IEEE standard is relatively thick (approximately 1 cm or 0.4 inches in diameter), and may be up to 50 meters (164 feet) long. There is no minimum length standard for a transceiver cable, and external transceivers were made that were small enough to fit directly onto the 15-pin AUI connector of the Ethernet interface, dispensing with the need for a transceiver cable.

"Office-grade" transceiver cables (shown at the bottom of Figure C-4) are thinner and more flexible than the standard cables. The thinner wires used in office-grade transceiver cables also have higher signal loss than standard cables, which limits the length of these cables. The maximum allowable length for an office-grade transceiver cable, which has four times the amount of signal attenuation as a standard cable, is 12.5 meters (41 feet).

Theoretically, you could connect several transceiver cables together to make up a single longer cable. However, this is not a good idea, because the sliding latch connectors may not hold the cable ends together very well, and you could easily end up with intermittent connections.

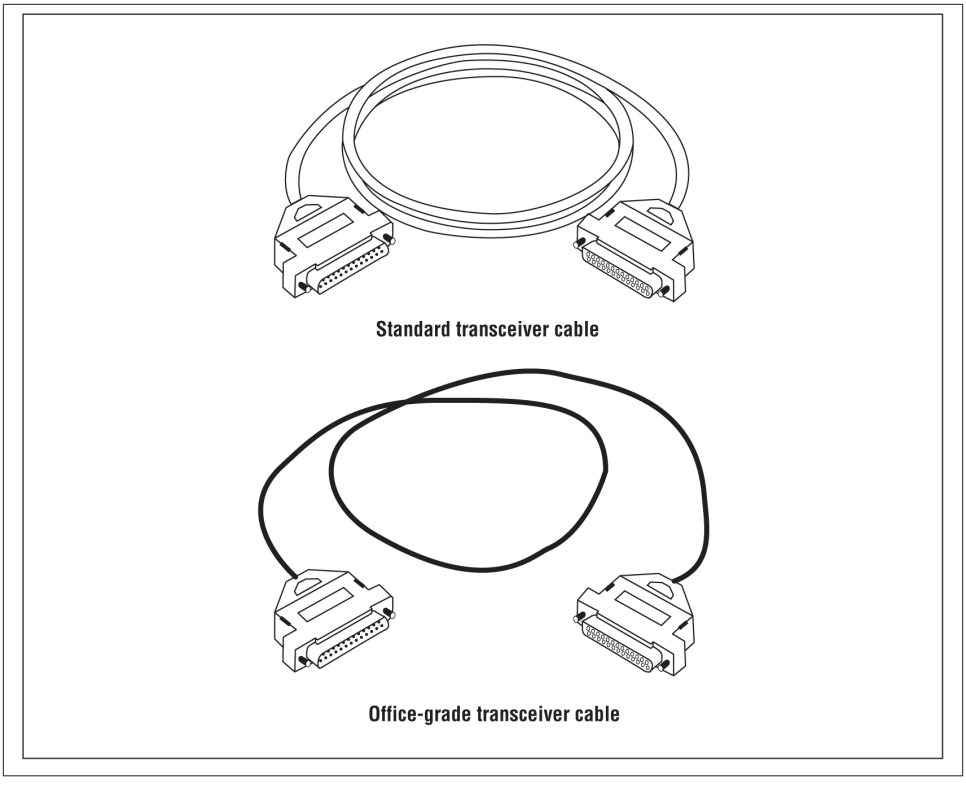


Figure C-4. Standard and office-grade AUI transceiver cables

The Medium Attachment Unit

The next component shown in the DTE 2 connection in Figure C-1 is the *medium attachment unit* (MAU), more commonly known as a *transceiver*. The transceiver gets its name because it *trans*mits and re*ceive*s signals on the physical medium. The AUI transceiver is the link between the different types of electrical signaling used over the media systems, and the signals that are sent through the AUI interface to the Ethernet interface in the station. Each 10 Mb/s media system has a specific transceiver designed to perform the type of signaling used for that medium. There are coaxial, twisted-pair, and fiber optic transceivers, each equipped with the components it takes to send and receive signals over that particular medium.

The external AUI transceiver is a small box, typically a few inches on each side. There is no specified shape; some are long and thin, and some almost square. The transceiver electronics typically receive their power over the transceiver cable from the Ethernet interface in the station. According to the standard, an AUI transceiver can draw up to 500 milliamps (0.5 amps) of current.

The transceiver transmits signals from the Ethernet interface onto the medium, and receives signals from the medium that it sends to the Ethernet interface. The signals that transceivers send vary according to the type of medium in use. On the other hand, signals that travel between the transceiver and the Ethernet-equipped device over the AUI interface are the same, no matter which media type is in use. That's why you can make a connection between any 10 Mb/s media system and a 15-pin connector on an Ethernet device. The signaling over the 15-pin interface is the same for all transceivers; only the media signals are different.

Transceiver Jabber Protection

The jabber protection function senses when a broken Ethernet device has gone berserk and is continuously transmitting a signal—a condition known as *jabbering*. Jabbering causes a continual carrier sense on the channel, jamming the channel and preventing other stations from being able to use the network. If that happens, the jabber protection circuit enables a *jabber latch*, which will shut off the signal to the channel.

The transceiver specification allows two methods of resetting the jabber latch: power-cycling, or automatically restoring operation a half second after the jabbering transmission ceases. In some very old transceivers, the jabber latch would not reset until the transceiver was power-cycled, which required the network administrator to disconnect and reconnect the transceiver cable to get the transceiver to work again. More modern transceivers were built using a single chip design that automatically came out of jabber latch mode once an overlong transmission had ceased.

The SQE Test Signal

The earliest Ethernet standard, DIX V1.0, did not include a signal for testing the operation of the collision detection system. However, in the DIX V2.0 specifications, the AUI transceiver was provided with a new signal called the *collision presence test* (CPT). The name for the collision signal changed to *signal quality error* (SQE) in the IEEE 802.3 standard, so the CPT signal was changed to SQE Test. The purpose of the SQE Test signal is to test the collision detection electronics of the transceiver, and to let the Ethernet interface know that the collision detection circuits and signal paths are working correctly. The SQE Test signal is also nicknamed the *heartbeat*, because it occurs regularly after each frame transmission.

When you install an external AUI transceiver on your Ethernet system, it is extremely important to correctly configure the SQE Test signal. The SQE Test signal *must be disabled* if the transceiver is attached to a repeater hub. For all other devices that may be attached to an external 10 Mb/s transceiver, the standard recommends that the SQE Test signal be enabled.

You may find that some vendors do not label things correctly, which can lead to some confusion. For example, you may find that the switch on the transceiver for enabling the SQE Test signal will be labeled "SQE" instead of "SQE Test." Because "SQE" is the name of the actual collision signal, the last thing you'd want to do is disable this signal in Ethernet. Nonetheless, this confusion of terms is very widespread.

Operation of SQE Test

The way the SQE Test signal works is simple. After every frame is sent, the transceiver waits a few bit times and then sends a short burst (about 10 bit times) of the collision presence signal. This signal is sent over the collision signal wires of the transceiver cable back to the Ethernet interface. This tests both the collision detection electronics and the signal paths.

Figure C-5 shows that the SQE Test signal travels from the external transceiver to the Ethernet interface. The interface in the computer receives the SQE Test signal on the collision signal wires of the transceiver cable after every frame transmission made by the interface.

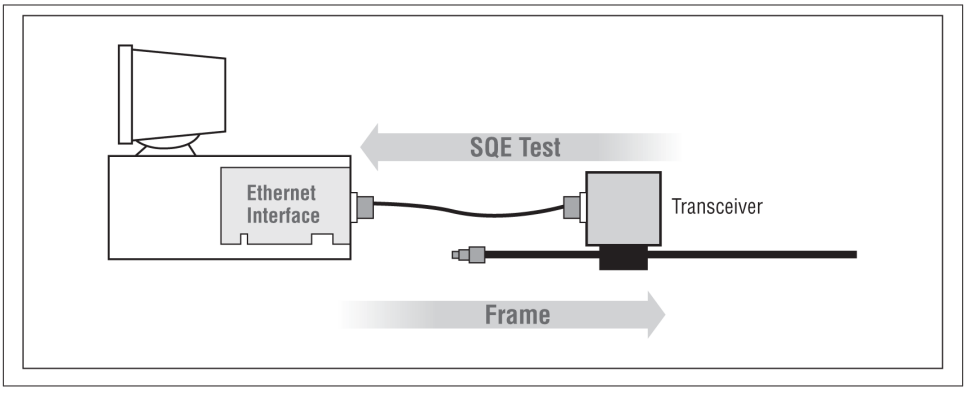


Figure C-5. SQE Test operation

Here are four important things to know about the SQE Test signal:

- The SQE Test signal is not sent out onto the network segment; it is only sent between the transceiver and the Ethernet interface as a test of the collision detection circuits.
- The SQE Test signal does not delay frame transmissions. Because the SQE Test pulse occurs during the interframe gap, no time is lost due to SQE Test signals. An Ethernet interface can send frames as fast as possible while also receiving SQE Test signals between every frame transmission.
- Although the SQE Test signal uses a short burst of the same signals used for a collision, SQE Test signals are not interpreted as a collision by the station. The

timing of the SQE Test pulse allows the station to differentiate the SQE Test signal from a real collision signal.

- The SQE Test signal *must* be disabled on an external transceiver connected to a repeater hub.

When the SQE Test signal was first introduced in the early 1980s, transceivers could be purchased without SQE Test, or with switch-selectable SQE Test (allowing you to turn it off). Eventually, all 10 Mb/s AUI-equipped transceivers came with a jumper or a switch that allowed the SQE Test signal to be disabled.

Ethernet Stations and SQE Test

For normal stations (DTEs) attached to a network segment with external AUI transceivers, the standard recommends that the SQE Test signal be enabled on the external transceivers. That's because the absence of a SQE Test signal after a frame transmission can alert the Ethernet interface that there may be a problem with the collision detection circuits. The problem could be caused by something simple, like a loose transceiver cable. On the other hand, it could be indicative of a more serious problem, like a failed collision detection circuit in the external transceiver.

Without a correctly functioning collision detection system, the Ethernet interface might ignore collisions on the network and transmit at incorrect times. While rare, this kind of failure can be difficult to debug. Ideally, network management software could alert you if it detected a problem or the absence of a SQE Test signal after a frame transmission.

However, it can be difficult to derive any benefit from the SQE Test signal in the real world. Most Ethernet interface software is designed not to make a fuss if the SQE Test signal is missing, mainly because the SQE Test is an optional signal on external transceivers. Many vendors took the approach of silently logging the presence or absence of a SQE Test signal in a software counter somewhere. This avoided support requests from people wondering what the error message about the SQE Test might mean.

There are other possible side effects of enabling SQE Test for external transceiver connections to normal stations. For example, the SQE Test signal can cause the collision presence light to flash on some transceivers and interfaces equipped with troubleshooting lights. That's because the SQE Test pulse is sent over the same pair of collision presence wires in the AUI cable as a real collision signal. This can cause the troubleshooting light to flash for both real collisions and SQE Test signals. Therefore, if you enable SQE Test—as recommended by the standard for all normal computers—you may need to ignore the effect that the SQE Test signal has on any collision presence lights on your network hardware.

The AUI Port Concentrator

Although it isn't described in the Ethernet standard, the AUI port concentrator, also called a *port multiplier*, *transceiver multiplexor*, or *fan out unit*, was widely used in older 10 Mb/s Ethernet systems. A basic concentrator is shown in Figure C-6.

The port concentrator was originally developed by the Digital Equipment Corporation (DEC) and called the DELNI (for Digital Ethernet Local Network Interconnect). Port concentrators sold by other vendors were often referred to as DELNIs or called "DELNI-like" devices. The port concentrator was developed when thick coaxial Ethernet was the only media type available, and network designers faced a problem when it came to connecting a set of machines clustered together in a small space.

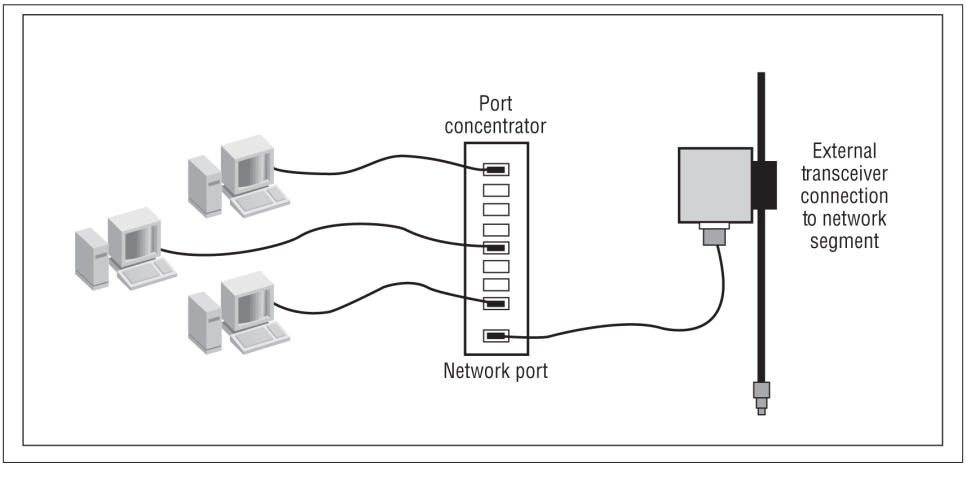


Figure C-6. Port concentrator

The problem arose because the thick Ethernet standard requires that each transceiver attachment be separated by at least 2.5 meters (8.2 feet) of thick coaxial cable from the next transceiver attachment. When you needed to connect a cluster of machines to the network, you had to coil up enough thick Ethernet coax to provide sufficient cable to meet the 2.5 m transceiver spacing requirement.

By providing several (usually eight) AUI ports in a single device, the DELNI makes it easier to connect groups of computers to an Ethernet. The eight computers are attached to the 15-pin male AUI connectors on the port concentrator. The concentrator has its own 15-pin female "network" port, which provides a way to connect the concentrator to a network segment using an external transceiver. In effect, all eight computers are sharing the single external transceiver connection to the network segment. The eight computers are not penalized for sharing a single transceiver connection, because only one computer can transmit at any given time on a half-duplex coaxial Ethernet system.

It's important to note that a port concentrator is not a repeater and does not include any signal retiming or regeneration circuits. In terms of the timing delay budget for a standard Ethernet, the concentrator unit sits in the transceiver cable path between the Ethernet segment and the station. The added signal delay and other effects contributed by the electronics inside the concentrator are not accounted for in the official configuration guidelines provided in the IEEE standard. These effects on the signal can vary depending on which vendor built the concentrator. In addition, because the concentrator is not described in the standard, a network system using concentrators cannot be verified using the IEEE configuration guidelines.

Port Concentrator Guidelines

To account for the added delay imposed by the port concentrator's electronics, some vendors state that you must use a shorter transceiver cable for connecting a station to the port concentrator. What they are saying is that you cannot use a full-length 50 m transceiver cable because the port concentrator has internal delays of its own that need to be accounted for. One vendor noted that the electronics in its port concentrators can add a signal delay equal to about 10 m of transceiver cable. Therefore, calculation of the total length of transceiver cable from the station to the live network must also include 10 m of port concentrator cable equivalence.

An easy way to deal with all of this is to simply add the 10 m cable equivalency of the port concentrator to the length of transceiver cable used to attach the port concentrator to the Ethernet segment. This figure provides a baseline transceiver cable length for that particular port concentrator installation. When you connect a station to the port concentrator, you need to add the transceiver cable length from the station to the port concentrator to the baseline figure to get the total transceiver cable length. The total length of the station transceiver cable plus the port concentrator baseline figure must not exceed 50 meters. It would no doubt be safer to make a total of 40 m your maximum value when using a port concentrator.

Calculating port concentrator cable length

If you use office-grade transceiver cables, you need to remember that the office cable has its own cable equivalency, which is four times the delay of an equivalent length of standard transceiver cable.

For example, a station attached to a port concentrator with a 5 m office-grade transceiver cable is equivalent to an attachment with 20 m of standard-grade transceiver cable. That 20 m distance must be added to the port concentrator internal cable equivalency of 10 m. To make the correct calculation of transceiver cable length, you must also include the length of transceiver cable used to connect the port concentrator to the external transceiver on the Ethernet segment.

Port concentrators can cause signal distortion due to their placement in the frame transmission path between the stations attached to the port concentrator and the stations on the rest of the network. As a signal propagates through the Ethernet system, it is allowed to accumulate a certain amount of timing distortion, known as *jitter*. Each component in the system has a jitter budget that it must not exceed for the system to work correctly. For instance, the standard for the AUI cable includes a jitter budget of ± 1 ns. This means that the signal traveling through an AUI cable is allowed to shift up to one nanosecond in either direction in time from its original time base.

Port concentrators have a set of electronics in them that cause a certain amount of jitter, and it's hard to design a port concentrator that will cause 1 ns or less of jitter in a signal. Therefore, you may find that a port concentrator will end up adding more jitter than the standard AUI cable is allowed to add according to the IEEE specifications. The accumulation of jitter in the signal is another reason why vendors limit the number of port concentrators you may connect together.

Problems with Concentrators

Cable equivalency in port concentrators, as well as the extra delay in office-grade transceiver cables and the accumulation of jitter, can easily be overlooked, leading to problems with the network connection. If you end up with too much signal delay or too much signal jitter in the path between the station and the network connection, the operation of the network may be affected. It can be hard to predict where things will fail because various components in your network (e.g., transceivers and Ethernet interfaces), may perform differently in the presence of excessive jitter. Some interfaces may be able to pick out a signal, while others may simply fail to receive the frame.

Experience shows that problems with port concentrator connections can lead to failures in which the amount of Ethernet frame loss can become quite high. Large frames seem to suffer the highest loss rate, while smaller frames can often make it through a marginal port concentrator connection with a lower loss rate. Some frames do manage to get through, so the network may appear to function at first glance. However, when frames are lost, the network application must recover by resending them. The application software typically has a timeout based on some number of seconds of no response time, after which it will retransmit a frame. This is a slow process, and one reason why a poorly configured port concentrator connection can appear to the user as a very slow network.

Cascaded Port Concentrators

It is possible to plug the network port of one port concentrator box into a station port of another concentrator, producing a cascade of two or more port concentrators. Cascading port concentrators means that you are adding the port concentrator delays together because the signals from the stations attached to the second port concentrator must go through the first port concentrator to reach the network segment.

The added delay and the accumulation of signal jitter can cause problems, which is why some vendors warn against this topology when the port concentrators must be attached to a *live* network (defined as one or more normal network segments supporting normal station connections as well as the port concentrator connection). In other words, cascading two standalone port concentrator units together will probably work. However, connecting the cascaded concentrators to an external network cable that is also supporting normal station connections may cause signal timing problems. The problems will occur due to the accumulation of timing delay and jitter in the combination of signal paths consisting of the cascaded port concentrators and the segment(s) making up the external network segment.

SQE Test and the Port Concentrator

When using port concentrators, you need to be aware of the way that they deal with the SQE Test signal. A port concentrator connected to an external transceiver with the SQE Test signal enabled will pass the SQE Test signal received from the external transceiver through every concentrator port. As such, every station attached to the port concentrator will receive SQE Test signals.

In a normal station, this is usually no problem. However, if you have a repeater attached to a port concentrator, then you need to make sure that the repeater does not receive the SQE Test signal. You can do this by turning off the SQE Test signal on the external transceiver that connects the port concentrator to the rest of the network.

If you are running the port concentrator in standalone mode, you may find that the port concentrator generates its own SQE Test signal internally and sends it out on all ports. Again, this can cause problems if you have a repeater attached to one of the ports of the port concentrator.

The Medium Dependent Interface

The actual connection to the network medium (e.g., twisted-pair cable) is made by way of a component that the standard calls the *medium dependent interface* (MDI). In the real world, this is a piece of hardware used for making a direct physical connection to the network cable.

In Figure C-1, the MDI is an eight-pin connector, also referred to as an RJ45-style jack. The MDI is actually a part of the transceiver, and it provides the transceiver with a direct physical and electrical connection to the twisted-pair wires used to carry network signals in the 10 Mb/s twisted-pair media system.

In the case of thick coaxial Ethernet, the most commonly used MDI was a type of coaxial cable clamp that installed directly onto the cable. For fiber optic Ethernet, the MDI is a fiber optic connector.

The Media Independent Interface

The invention of the 100 Mb/s Fast Ethernet system was also the occasion for the development of a new attachment interface, called simply the *media independent interface* (MII). The MII can support operation at both 10 and 100 Mb/s.

Figure C-7 shows the same two stations as Figure C-1, but this time an MII is being used. The other major difference between this diagram and the one shown in Figure C-1 is that a transceiver with an MII is called a *physical layer device* (PHY) instead of a MAU. In essence, the MII is an updated and improved version of the original 10 Mb/s–only AUI. The MII may be embedded inside equipment, or may not be used at all. In this case, the transceiver is embedded in the computer. All the user sees is the twisted-pair connector, which connects the twisted-pair medium to the internal transceiver.

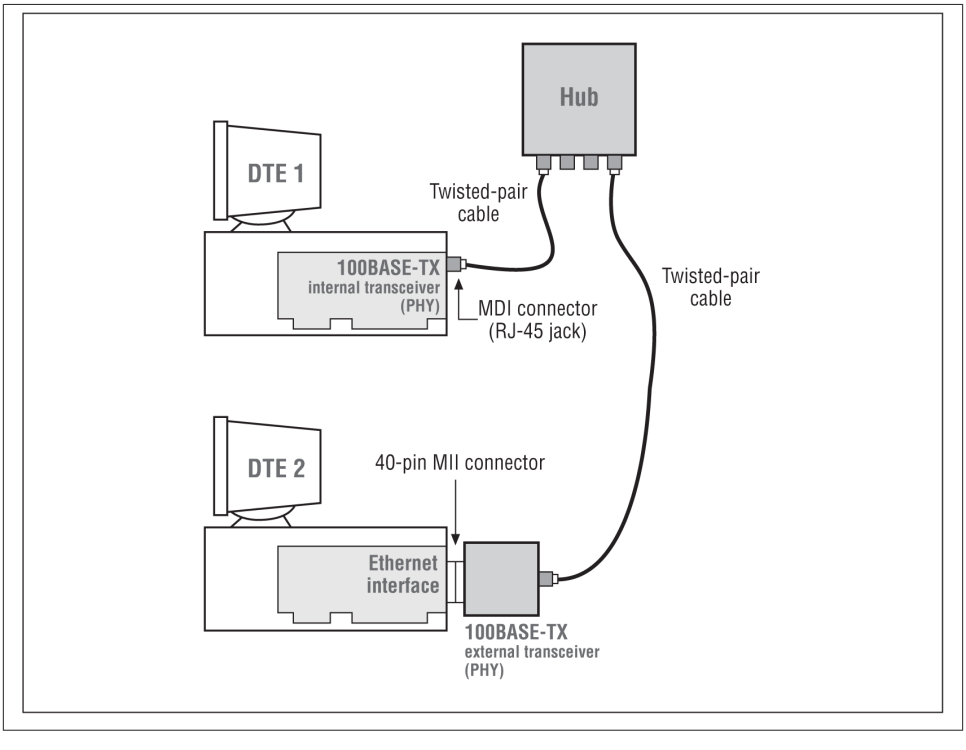


Figure C-7. MII connection for a 100 Mb/s Fast Ethernet system

An Ethernet device could also be provided with a 40-pin MII connector that allows connections to external transceivers, as shown attached to DTE 2 in Figure C-7. The external transceiver provides flexibility because you can provide either a twisted-pair

or fiber optic transceiver. This allows a connection to either a twisted-pair or fiber optic media types operating at 10 or 100 Mb/s speeds.

The MII is designed to make the signaling differences among the various media segments transparent to the Ethernet electronics inside the networked device. It does this by converting the signals the transceiver (PHY) receives from the various media segments into standardized digital format signals. The digital signals are then provided to the Ethernet electronics in the networked device over a 4-bit-wide data path. The same standard digital signals are provided to the Ethernet interface no matter what kind of signaling is used on the media system.

The MII Connector

The 40-pin MII connector and optional MII cable provide a path for the transmission of signals between an MII interface in the station and an external transceiver. The vast majority of external MII transceivers are designed for direct connection to the MII connector on the networked device and do not use an MII cable. Today, all twisted-pair connections are made directly to an eight-pin (RJ45) connector on the Ethernet device or switch port. No one uses external transceivers for twisted-pair connections anymore.

Figure C-8 shows two MII transceivers, one equipped with an MII cable (top), and the other (below) equipped with jack screws for direct connection to the mating screw locks on the DTE. The jack screws replaced the much-maligned slide latch mechanism used for the 15-pin AUI in the original 10 Mb/s Ethernet system. If the optional MII cable is used, the end of the MII cable is equipped with a 40-pin connector and a pair of jack screws that fasten into the mating screw locks on the networked device.

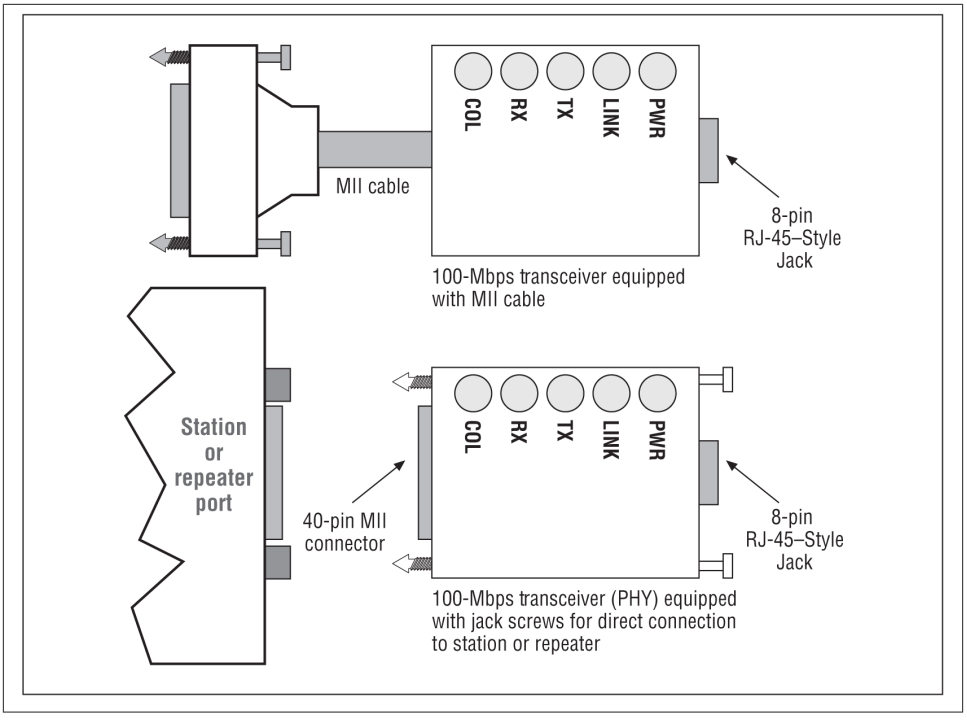


Figure C-8. MII connector and transceiver

MII connector signals

The signals provided on the MII connector are different from the ones found on the 15-pin AUI connector in the 10 Mb/s system.

The external 40-pin MII connector is small and the pins are densely packed, so care should be taken to avoid damaging the pins when connecting and disconnecting network components. Note that the MII pins can easily be bent and that the +5 volt pins on the lower row are right next to the ground pins.

If the +5 volt pins or ground pins bend and touch one another during installation, it is possible to blow a fuse in the network equipment, which will cause the MII port to stop working. The prudent network manager may wish to power off the equipment while connecting or disconnecting MIIs, just to be safe.

Figure C-9 shows the 40 pins of the MII connector, with the signals carried by the pins indicated. The MII defines a 4-bit-wide data path for transmit and receive data that is clocked at 25 MHz to provide a 100 Mb/s transfer speed, or 2.5 MHz for 10 Mb/s operation. According to the standard, the electronics attached to each MII connector

(male and female) should be designed to withstand connector insertion and removal while the power is on.

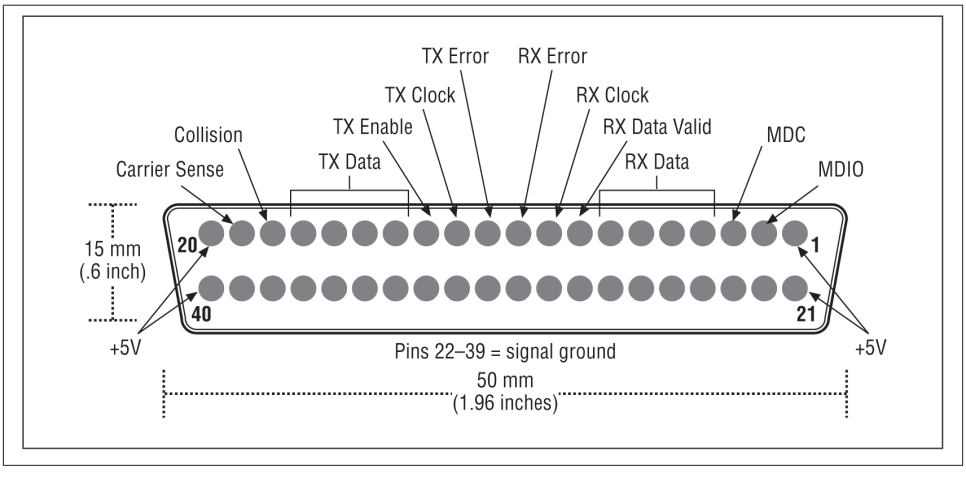


Figure C-9. MII signals

The MII provides for a set of control signals that make it possible for the Ethernet interface in the networked device to interact with the external transceiver to set and detect various modes of operation. This management interface can be used to place the transceiver into loopback mode for testing, to enable full-duplex operation if the transceiver supports it, to select the transceiver speed if the transceiver supports dual-speed mode, and so on. The MII signals are as follows:

+5 volts
: Pins 1, 20, 21, and 40 are used to carry +5 volts at a maximum current of 750 milliamps, or 0.75 amps.

Signal Ground
: Pins 22 through 39 carry signal ground wires.

Management Data I/O
: Pin 2 provides the Management Data Input/Output signal, a bi-directional signal used to carry serial data representing control and status information between the transceiver and the DTE. The management interface provides various functions, including resetting the transceiver, setting the transceiver to full-duplex mode, and enabling a test of the electronics and signal paths involved in the collision signal.

Management Data Clock
: Pin 3 provides the Management Data Clock, which is used as a timing reference for serial data sent on the management data interface.

RX Data
: Pins 4, 5, 6, and 7 provide the 4-bit Receive Data path from the transceiver to the DTE.[2]

RX Data Valid
: Pin 8 provides the Receive Data Valid signal, which is generated by the transceiver while receiving a valid frame.

RX Clock
: Pin 9 provides the Receive Clock, which runs at 25 MHz in 100 Mb/s Fast Ethernet systems and at 2.5 MHz in 10 Mb/s systems, to provide a timing reference for receive signals.

RX Error
: Pin 10 carries this signal, which is sent by the transceiver upon detection of received errors.

TX Error
: Pin 11 carries this signal, which can be used by a repeater to force the propagation of received errors. This signal may be used by a repeater under certain circumstances, but is never used by a station.

TX Clock
: Pin 12 provides the Transmit Clock, which runs continuously at a frequency of 25 MHz for 100 Mb/s Fast Ethernet systems and at 2.5 MHz for 10 Mb/s systems. The purpose of this signal is to provide a timing reference for transmit signals.

TX Enable
: Pin 13 carries the Transmit Enable signal from the DTE to the transceiver to signal the transceiver that transmit data is being sent.

TX Data
: Pins 14, 15, 16, and 17 provide the 4-bit-wide Transmit Data path from the DTE to the transceiver.

Collision
: Pin 18 carries this signal from the transceiver. It indicates a collision detected on the network segment. If a transceiver is in full-duplex mode, then this signal is undefined by the standard, and the collision light on the transceiver may glow steadily or erratically when full-duplex mode is enabled. See the instructions that came with the transceiver you are using for details.

2. A 4-bit chunk of data is also called a *nibble*, to distinguish it from an 8-bit byte.

Carrier Sense
: Pin 19 carries this signal, which indicates activity on the network segment from the transceiver to the DTE.

MII Transceivers and Cables

The PHY shown in Figure C-7 performs approximately the same function as a transceiver in the 10 Mb/s Ethernet system. However, unlike the original 10 Mb/s MAU, the PHY also performs the medium system signal encoding and decoding. In addition, an MII transceiver can be automatically configured to operate in full- or half-duplex mode and can operate at either 10 or 100 Mb/s.

The transceiver may be a set of integrated circuits inside the Ethernet port of a network device and therefore invisible to the user, or it may be a small box like the external transceiver used in 10 Mb/s Ethernet. An external MII transceiver is equipped with a 40-pin MII plug designed for direct connection to the 40-pin MII jack on the networked device, as shown in Figure C-8. This connection may include a short MII transceiver cable, although these were rarely used.

According to the standard, an MII cable consists of 20 twisted pairs with a total of 40 wires. The twisted-pair cable also has a 40-pin plug on one end equipped with male jack screws that connect to mating female screw locks. The cable can be a maximum of 0.5 meters in length (approximately 19.6 inches). However, the vast majority of external transceivers attach directly to the MII connector on the device with no intervening cable.

MII jabber protection

An MII transceiver operated at 10 Mb/s has a jabber protection function that provides a jabber latch similar to the one in the AUI transceiver described earlier in this chapter. In the 100 Mb/s Fast Ethernet system, the jabber protection feature was moved to the Fast Ethernet repeater ports. This change was possible because all Fast Ethernet segments are link segments and must be connected to a repeater hub for communication with other stations in half-duplex mode.

Moving the circuitry for jabber protection to the repeater hub offloads that requirement from the transceiver and provides the same level of protection for the network channel. Therefore, transceivers operating an 100 Mb/s Fast Ethernets do not provide the jabber latch function. Instead, each Fast Ethernet repeater port monitors the channel for long transmissions and shuts the port down if the carrier signal persists for anywhere from 40,000 to 75,000 bit times.

MII SQE Test

The SQE Test signal is provided on AUI-based equipment to test the integrity of the collision detection electronics and signal paths. However, there is no SQE Test signal

provided in the MII. SQE Test can be removed from the MII because all media systems connected to an MII are link segments. Collisions are detected on link segments by the simultaneous occurrence of data on the receive and transmit data circuits. As such, the link monitor function in MII transceivers ensures that the receive data circuits are working correctly.

Additionally, the MII provides a loopback test of the collision detect signal paths from an external transceiver to the Ethernet device. Taken together, this provides a complete check of collision detect signal paths, making the SQE Test signal unnecessary for the MII.

Glossary

4B/5B
: A block encoding scheme used to send Fast Ethernet data. In this signal encoding scheme, 4 bits of data are turned into 5-bit code symbols for transmission over the media system.

4D-PAM5
: A block encoding scheme used for 1000BASE-T twisted-pair Gigabit Ethernet that transmits signals over four wire pairs. This coding scheme translates an 8-bit byte of data into a simultaneous transmission of four code symbols (4D) that are sent over the media system as 5-level Pulse Amplitude Modulated (PAM5) signals.

50-pin connector
: A connector that is sometimes used on 10BASE-T hubs as an alternate twisted-pair segment connection method. The 50-pin connector is used to connect 25-pair cables used in telephone wiring systems, which are typically rated to meet Category 3 specifications. Commonly referred to as a Telco, CHAMP, or "blue ribbon" connector.

8-pin connector
: A twisted-pair connector that closely resembles the RJ45 connector used in US telephone systems, but has significantly better electrical characteristics than typical telephone-grade RJ45 connectors.

8B6T
: A block encoding scheme used in 100BASE-T4 systems, based on translating 8-bit (binary) data patterns into 6-bit code symbols that are transmitted as 3-level (ternary) signals.

8B/10B
: A block encoding scheme used in 1000BASE-X Gigabit Ethernet systems, in which 8-bit bytes of data are translated into 10-bit code symbols.

10BASE-T
: 10 Mb/s Ethernet system based on Manchester signal encoding transmitted over Category 3 or better twisted-pair cable.

10BASE2
: 10 Mb/s Ethernet system based on Manchester signal encoding transmitted over thin coaxial cable. Also called *Thin Wire* and *Cheapernet*.

10BASE5
: 10 Mb/s Ethernet system based on Manchester signal encoding transmitted over thick coaxial cable. Also called *Thick Net*.

10BASE-F
: 10 Mb/s Ethernet system based on Manchester signal encoding transmitted over fiber optic cable.

100BASE-FX
: 100 Mb/s Fast Ethernet system based on 4B/5B signal encoding transmitted over fiber optic cable.

100BASE-T
: A term that is confusingly used both for the entire 100 Mb/s Fast Ethernet system and for the 100 Mb/s twisted-pair system by itself.

100BASE-T2
: 100 Mb/s Fast Ethernet system designed to use two twisted pairs in a Category 3 twisted-pair cable.

100BASE-T4
: 100 Mb/s Fast Ethernet system based on 8B6T signal encoding transmitted over four twisted pairs in a Category 3 twisted-pair cable.

100BASE-TX
: 100 Mb/s Fast Ethernet system based on 4B/5B signal encoding transmitted over two twisted pairs in a Category 5 twisted-pair cable.

100BASE-X
: A term used when referring to any Fast Ethernet media system based on 4B/5B block encoding. Includes 100BASE-TX and 100BASE-FX media systems.

802.1
: IEEE Working Group for High-Level Interfaces, Network Management, Internetworking (including bridges), and other issues common across LAN technologies.

802.2
: IEEE Working Group for Logical Link Control (LLC).

802.3
: IEEE Working Group for Ethernet LANs.

1000BASE-CX
: 1000 Mb/s Gigabit Ethernet system based on 8B/10B block encoding that is transmitted over copper cable.

1000BASE-LX
: 1000 Mb/s Gigabit Ethernet system based on 8B/10B block encoding that is transmitted using long-wavelength laser transmitters and fiber optic cable.

1000BASE-SX
: 1000 Mb/s Gigabit Ethernet system based on 8B/10B block encoding that is transmitted using short-wavelength laser transmitters and fiber optic cable.

1000BASE-T
: 1000 Mb/s Gigabit Ethernet system based on 4D-PAM5 block encoding that is transmitted over twisted-pair cabling.

1000BASE-X
: A term used when referring to any 1000 Mb/s (Gigabit) media system based on the 8B/10B encoding scheme used in Fibre Channel. Includes 1000BASE-CX, 1000BASE-LX, and 1000BASE-SX.

10GBASE-CX4
: A short-reach copper cable media system initially defined in the 802.3ak supplement to the standard. The CX4 specifications, adopted into the standard in 2004 as Clause 54, define a media system based on twinaxial cables and equipped with 16-pin InfiniBand connectors.

10GBASE-LRM
: The long-reach multimode (LRM) media type is designed to operate over multimode fiber optic cables that meet the 10GBASE-S fiber optic specifications, using 1,310 nm laser light sources.

10GBASE-LX4
: This media type is designed to operate over both single-mode and multimode fiber optic cables using four separate laser light sources.

10GBASE-SR
: Designed for short-reach applications, this media type operates over a single pair of multimode fiber optic cables that meet the 10GBASE-S fiber optic specifications.

10GBASE-T
: 10 Gigabit Ethernet system transmitted over twisted-pair cable.

10GSFP+Cu
: This media type is not specified in the 802.3 standard; the shorthand identifier was invented by the vendors who developed this short reach direct attach copper cable, equipped with SFP+ connectors.

40GBASE-CR4
: The 40GBASE-CR4 short reach copper segment is defined in Clause 85 of the standard. It specifies a media system based on four lanes of PCS data carried over four twinaxial cables, equipped with QSFP+ connectors.

40GBASE-LR4
: The 40GBASE-LR4 long reach media system is designed to operate over single-mode fiber optic cables.

40GBASE-SR4
: The 40GBASE-SR4 short-range media type is designed to operate over multimode fiber optic cables.

40GBASE-T
: 40 Gigabit Ethernet system transmitted over twisted-pair cable.

Address
: A means of uniquely identifying a device on a network.

ANSI
: American National Standards Institute. The coordinating body for many voluntary standards groups within the United States, and the U.S. representative to the International Organization for Standardization (ISO).

ARP
: Address Resolution Protocol. A protocol used to discover a destination host's hardware (MAC) address when given the host's IP address.

ASIC
: Application Specific Integrated Circuit. An integrated circuit chip specifically designed for a given application. ASICs allow vendors to develop high-performance network devices (such as switches) with more capabilities at lower cost.

ASN.1
: Abstract Syntax Notation-1. An ISO standard for machine-independent data formatting, used as the data formatting standard for SNMP MIBs.

Attenuation
: The decreasing power of a transmitted signal as it travels along a cable. The longer a cable is, the greater the signal attenuation will be. This loss is expressed in decibels (dB).

AUI
: Attachment Unit Interface. The 15-pin signal interface defined in the original Ethernet standard that carries signals between a station and an outboard transceiver.

AUI cable
: Also known as a transceiver cable, the AUI cable connects a station to an outboard transceiver.

AUI connector
: The 15-pin AUI connector on a station, cable, or outboard transceiver that allows these devices to be interconnected.

Auto-Negotiation
: A protocol defined in the Ethernet standard that allows devices at either end of a link segment to advertise and negotiate modes of operation such as the speed of the link. Other modes that can be negotiated include full- or half-duplex operation and support for Ethernet flow control. If a device is equipped with Auto-Negotiation, it can determine the capabilities of the device at the other end of the link (the link partner) and select the highest common denominator of operational modes.

AWG
: American Wire Gauge. This is a U.S. standard set of wire conductor sizes. "Gauge" means the diameter of the wire; the higher the gauge number, the smaller the diameter and the thinner the wire. The gauge is measured in decimal fractions of an inch. For example, a 22 AWG wire has a diameter of 0.02534 inches.

Backbone
: A network used as a primary path for transporting traffic between network segments. A backbone network is often based on higher-capacity technology, to provide enough bandwidth to accommodate the traffic of all segments linked to the backbone.

Bandwidth
: The maximum capacity of a network channel, usually expressed in bits per second (bps). Commonly used Ethernet channels have a bandwidth of between 10 and 100 Gb/s.

Baud
: A unit of signaling speed. The speed in baud is the number of discrete signal events per second. If each signal event represents a single bit, then the baud rate is the same as the bit rate per second. If more than one signal event is required to transmit a bit of data, then the baud rate will be greater than the rate of bits per second.

Bit
: The smallest unit of data, either a zero or a one.

Bit error rate
: Also called Bit Error Ratio in the standard. A measure of data integrity, expressed as the ratio of received bits that are in error, relative to the amount of bits received. Often expressed as a negative power of 10. For example, the worst-case bit error rate for several 10 Mb/s Ethernet media varieties is 10^{-9} (a rate of one error in every 1 billion bits transmitted, on average).

Bit time
: The length of time required to transmit one bit of information.

Block encoding
: A system whereby a group of data bits are encoded into a larger set of code bits. The data stream is divided into a fixed number of bits per block. Each data block is translated into a set of code bits, also called code symbols. The expanded set of code symbols is used for control purposes, such as start-of-frame, end-of-frame, carrier extension, and error signaling.

BNC
: A bayonet locking connector used on 10BASE2 thin coaxial cable segments. The BNC designation is said to stand for *Bayonet Navy Connector*. However, it is also said to stand for *Bayonet Neill-Concelman*, after the names of two designers of coaxial connectors.

Bridge
: A device that connects two or more networks at the data link layer.

Broadcast address
: The multicast destination address of all ones, defined as the group of all stations on a network. The standard requires that every station must receive and act upon every Ethernet frame whose destination address is all ones.

Broadcast domain
: The set of all nodes connected in a network that will receive each other's broadcast frames. All Ethernet segments connected with a Layer 2 bridge are in the same broadcast domain. Virtual LANs (VLANs) can be used to establish multiple broadcast domains in an Ethernet system based on switches.

Building entrance
: An area inside a building where cables enter and are connected to riser cables for signal distribution throughout the building.

Bus
: In general, an electrical transmission path for carrying information, usually serving as a shared connection for multiple devices. In LAN technology, a linear network topology, in which all computers are connected to a single cable.

Carrier sense
: In Ethernet, a method of detecting the presence of signal activity on a common channel.

Category 3
: Twisted-pair cable with a Category 3 rating has electrical characteristics suitable for carrying 10BASE-T and 100BASE-T4 signals. Category 3 cable is no longer recommended for use in building cabling systems.

Category 5
: Category 5 cable has electrical characteristics suitable for all twisted-pair Ethernet media systems, including 10BASE-T, 100BASE-TX, and 1000BASE-T. Category 5 and Category 5e cables are the preferred cable types for structured cabling systems.

Category 5e
: An enhanced version of Category 5 cable, developed to improve certain cable characteristics important to Gigabit Ethernet operation. It is recommended that all new structured cabling systems be based on Category 5e cable.

Coaxial cable
: A cable with low susceptibility to interference. An outer conductor, also called a screen or shield, surrounds an inner conductor. The conductors are commonly separated by a solid plastic or foam plastic insulating material. Thick and thin coaxial cables are used for 10BASE5 and 10BASE2 Ethernet systems, respectively.

Collision
: A normal event on a half-duplex Ethernet system that indicates simultaneous channel access by two or more stations. A collision is automatically resolved by the Media Access Control (MAC) protocol.

Collision detection
: A method of detecting two or more simultaneous transmissions on a common signal channel.

Conditioned launch cable
: A special fiber optic patch cable that offsets the coupling (launch) of laser light from the center of a fiber optic cable. This avoids the phenomenon of differential mode delay, which can occur when laser light sources are coupled to multimode fiber optic cables.

CoS
: Class of Service. The IEEE 802.1Q standard provides an extra field in the Ethernet frame to hold both a VLAN identifier and CoS tags. The CoS tag values are defined in the IEEE 802.1p standard.

CRC
: Cyclic Redundancy Check. An error checking technique used to ensure the accuracy of transmitted data. The frame fields other than the preamble are used in the process of mathematically computing a checksum, which is placed in the frame check sequence (FCS) field of the frame as the frame is transmitted. The receiving station uses the same process to compute a checksum and compares this checksum to the contents of the received frame's FCS field. Identical checksums indicate that the frame fields were received correctly.

Crossover cable
: A twisted-pair patch cable wired in such a way as to route the transmit signals from one piece of equipment to the receive signals of another piece of equipment, and vice versa.

Crosstalk
: The unwanted transfer of a signal from one circuit to another. In twisted-pair cables, the unwanted transfer of signals from transmitting wires to other wires in the cable plant. The maximum level of crosstalk is measured at the end nearest the transmitter, leading to the term near-end crosstalk (NEXT).

CSMA/CD
: Carrier Sense Multiple Access with Collision Detection. The formal name for the Media Access Control (MAC) protocol used in Ethernet.

D connector
: A family of connectors including the 25-pin RS232 connector, the 15-pin AUI connector, and the 9-pin connector. The outline of such a connector seen end-on is roughly that of the letter "D."

Data link layer
: Layer 2 of the OSI reference model. This layer takes data from the network layer and passes it on to the physical layer. The data link layer is responsible for transmitting and receiving Ethernet frames.

DCE
: Data Communications Equipment. Any equipment that connects to data terminal equipment (DTE) to allow data transmissions between DTEs.

DIW
: Direct Inside Wire. Twisted-pair cabling used inside a building, which usually contains four pairs of wires within the cable.

Drop cable
: The connection (drop) between a network device and an outlet. In the original Ethernet system, the transceiver cable was sometimes called a drop cable. In twisted-pair Ethernet systems the patch cable may also be called a drop cable.

DTE
: Data Terminal Equipment. Any piece of equipment at which a communications path begins or ends; that is, the data station (computer) serving as the data source, destination, or both, for the purpose of sending or receiving data on a network.

Encoding
: A means of combining clock and data information into a self-synchronizing stream of signals.

Error detection
: A method that detects errors in received data by examining cyclic redundancy checks (CRCs) or using other techniques.

Ethernet
: A popular local area networking (LAN) technology first standardized by DEC, Intel, and Xerox (DIX) and subsequently standardized by the IEEE.

Fast Ethernet
: A version of Ethernet that operates at 100 Mb/s.

Fast link pulse
: A link pulse that encodes information used in the Auto-Negotiation protocol. Fast link pulses consist of bursts of the normal link pulses used in 10BASE-T.

FDDI
: Fiber Distributed Data Interface. An ANSI standard (ANSI X3T12) for a 100 Mb/s token passing network (Token Ring) based on fiber optic and twisted-pair cable.

Fiber optic cable
: A cable with a glass or plastic filament that transmits digital signals in the form of light pulses.

Filter rate
: The maximum number of frames that a switch can continuously receive, inspect, and make a forwarding decision on.

Flow control
: The process of controlling data transmission at the sender to avoid overfilling buffers and loss of data at the receiver.

FOIRL
: Fiber Optic Inter-Repeater Link. An early version of fiber optic link segment defined in the IEEE 802.3c standard.

Forwarding
: The process of moving frames from one port to another in a switch.

Forwarding rate
: The maximum number of frames that can be forwarded by a switch, assuming no con-

gestion on the network to which the output port is connected.

Frame
: The fundamental unit of transmission at the data link layer of LAN operation.

Full-duplex media
: A signal transmission path that can support simultaneous data transmission and reception.

Full-duplex mode
: A communications method that allows a device to simultaneously send and receive data.

Gigabit Ethernet
: A version of Ethernet that operates at 1 billion (1,000,000,000) bits per second.

GMII
: Gigabit Media Independent Interface. Unlike the AUI or MII, the GMII is not a physical interface. Instead, the GMII is a logical interface used in the standard to define the set of signals that flow between Gigabit Ethernet transceiver chips and controller chips inside Gigabit Ethernet ports.

Half-duplex mode
: A communications method in which a device may either send or receive data at a given moment, but not both.

Heartbeat
: See *SQE test*.

Hub
: A device at the center of a star topology network. A hub device may be a repeater, bridge, switch, router, or any combination of these.

IEEE
: Institute for Electronics and Electrical Engineers. A professional organization and standards body. IEEE Project 802 is the group within the IEEE that is responsible for LAN technology standards.

IETF
: Internet Engineering Task Force. The technical group that sets standards for the community of people using TCP/IP network protocols.

Impedance
: A measure of opposition to the flow of a current at a particular frequency, measured in ohms.

Interframe gap
: An idle time between frames, also called the *interpacket gap*.

Internet
: A worldwide collection of networks based on the use of TCP/IP network protocols.

Intranet
: A collection of networks supporting a single site or corporate entity, linked at the network layer of operation using routers.

Jabber
: The act of continuously sending data. A jabbering station is one whose circuitry or logic has failed, and which has locked up a network channel with its incessant transmissions.

Jabber latch
: A protective circuit in Ethernet transceivers or repeater hubs that interrupts an overlong transmission.

Jitter
: Also called *phase jitter*, timing distortion, or intersymbol interference. The slight movement of a transmission signal in time or phase that can introduce errors and loss of synchronization. The amount of jitter will increase with longer cables, cables with higher attenuation, and signals at higher data rates.

Late collision
: A failure of the network in which the collision indication arrives too late in the frame transmission to be automatically dealt with by the Media Access Control (MAC) protocol. The frame being transmitted will be dropped, requiring that the application detect and retransmit the lost frame, which may result in greatly reduced throughput. Late collisions may be caused by a mis-

match in duplex settings at each end of a link. Another cause is excessive levels of signal crosstalk in a twisted-pair cabling system.

Latency
: A measure of the delay experienced in a system. In Ethernet switches, latency is the time required to forward a packet from the input (ingress) port to the output (egress) port.

LACP
: Link Aggregation Control Protocol. The IEEE 802.1AX Link Aggregation standard allows multiple parallel Ethernet links to be grouped together, functioning as a single "virtual" channel. A given packet flow over the channel is limited to a single link in the channel; therefore, single packet flows cannot exceed the speed of the individual links. However, multiple packet flows will be distributed across the multiple links in the channel, resulting in an aggregate throughput for multiple flows that is the sum of the speeds of the individual links in the group. Link aggregation was first defined in the IEEE 802.3ad standard, and later moved to become IEEE 802.1AX.

Link integrity test
: On link segments, a test that checks for the presence of link test pulses or link signaling activity. This test verifies that the link is connected correctly and that signals are being received correctly.

Link layer
: See *Data link layer*.

Link light
: An optional status light on a transceiver or interface card that indicates the status of the link integrity test. If this light is lit on both ends of the link, it indicates that the link is passing the link integrity test.

Link pulse
: A test pulse sent between transceivers on a 10BASE-T link segment during periods of no traffic, to test the signal integrity of the link.

Link segment
: Defined in the IEEE 802.3 specifications as a point-to-point segment that connects two —and only two—devices.

LLC
: Logical Link Control. A standardized protocol and service interface provided at the data link layer and independent of any specific LAN technology. Specified in the IEEE 802.2 standard.

Mbaud
: 1 million baud. See *Baud*.

MAC
: Media Access Control. A protocol defining a set of mechanisms, operating at the data link layer of a local area network. The MAC protocol is used to manage access to the communications channel.

MAC address
: The 48-bit address used in Ethernet to identify a station interface.

Manchester encoding scheme
: Signal encoding method used in all 10 Mb/s Ethernet media systems, including 10BASE2, 10BASE5, 10BASE-F, and 10BASE-T. Each bit of information is converted into a "bit symbol" that is divided into two halves. During the first half the signal being sent is the complement of the data bit being encoded. During the second half the signal is identical to the data symbol. This provides a signal transition in each bit symbol sent, which is used as a clock signal for synchronization by the receiving device.

MAU
: Medium Attachment Unit. The IEEE 802.3 name for the device called a transceiver in the original DIX Ethernet standard. The MAU provides the physical and electrical interface between an Ethernet device and the media system to which it is connected.

MDI
: Medium Dependent Interface. The name for the connector used to make a physical and electrical connection between a trans-

ceiver and a media segment. The 8-pin RJ45-style connector is the MDI for the 10BASE-T, 100BASE-TX, 100BASE-T4, and 1000BASE-T media systems.

MDI-X
: An MDI port on a hub that has an internal crossover signal. This means that a "straight-through" patch cable can be used to connect a station to this port, because the signal crossover is performed inside the port.

MIB
: Management Information Base. A list of manageable objects (counters, etc.) for a given device; used by management applications.

MIC
: Media Interface Connector. Specified for use in the FDDI LAN system to make a connection to a pair of fiber optic cables. May also be used in the 100BASE-FX media system; however, the duplex SC connector is listed in the specifications as the preferred connector for 100BASE-FX.

MII
: Media Independent Interface. Similar to the AUI, but designed to support both 10 and 100 Mb/s, an MII provides a 40-pin connection to outboard transceivers (also called PHY devices). Used to attach 802.3 interfaces to a variety of physical media systems.

Mixing segment
: Defined in the IEEE 802.3 specifications as a segment that may have more than two MDI connections. Coaxial Ethernet segments are mixing segments.

MSTP
: Multiple Spanning Tree Protocol. Originally defined in the IEEE 802.1s supplement to the IEEE 802.1Q standard. This version of spanning tree adds the facility for switches supporting VLANs to use multiple spanning trees, providing for traffic belonging to different VLANs to flow over different paths within the network. MSTP is an optional Spanning Tree Protocol that is supported in some switches.

Multicast address
: Allows a single Ethernet frame to be received by a group of stations. If the first bit of the destination address transmitted on the Ethernet channel is a one (1), then the address is a multicast address.

N connector
: A coaxial cable connector used for 10BASE5 thick coax segments. The connector is named after its developer, Paul Neill.

Network layer
: Layer 3 of the OSI reference model. The layer at which routing based on high-level network protocols occurs.

NIC
: Network Interface Card. Also called an adapter or interface card, this is the set of electronics that provides a connection between a computer and a LAN.

Octet
: Eight bits (also called a byte).

OSI
: Open Systems Interconnection. A seven-layer reference model for networks, developed by the International Organization for Standardization (ISO). The OSI reference model is a formal method for describing the interlocking sets of networking hardware and software used to deliver network services.

OUI
: Organizationally Unique Identifier. A 24-bit value assigned to an organization by the IEEE. Ethernet vendors use the 24-bit OUIs they receive from the IEEE in the process of creating unique 48-bit Ethernet addresses. Each Ethernet device a vendor builds is provided with a unique 48-bit address, whose first 24 bits are composed of the vendor's OUI.

Packet
: A unit of data exchanged at the network layer (Layer 3 of the OSI reference model).

PAM5x5
: A signal encoding scheme used in the 100BASE-T2 media system.

Patch cable
: A twisted-pair or fiber optic jumper cable used to make a connection between a network interface on a station or network port on a hub and a media segment, or to directly connect stations and hub ports together.

Phantom collision
: A false collision detect signal. In twisted-pair Ethernet systems, a phantom collision can be caused by excessive signal crosstalk. Collisions are detected on twisted-pair segments by the simultaneous presence of signals on the transmit and receive signal pairs. Excessive signal crosstalk on a twisted-pair segment can cause signals to simultaneously appear on both the transmit and receive signal pairs, which triggers a false, or phantom, collision indication to the transmitting interface.

Phase jitter
: See *Jitter*.

PHY
: Physical Layer Device. According to the 802.3 standard: "The Physical Layer encodes frames for transmission and decodes received frames with the modulation specified for the speed of operation,transmission medium and supported link length. Other specified capabilities include: control and management protocols, and the provision of power over selected twisted pair PHY types."[3]

Physical address
: The 48-bit MAC address assigned to a station interface, identifying that station on the network.

Physical layer
: The first layer in the OSI seven-layer reference model. This layer is responsible for physical signaling, including the connectors, timing, voltages, and related issues.

Plenum cable
: A cable that is rated as having adequate fire resistance and satisfactorily low smoke-producing characteristics for use in plenums (air-handling spaces). Air-handling spaces are often located below machine room floors, or above suspended ceilings.

Point-to-point topology
: A network system composed of point-to-point links. Each point-to-point link connects two and only two devices, one at each end.

Port
: A connection point for a cable. Repeater hubs and switches typically provide multiple ports for connecting Ethernet devices.

Promiscuous mode
: A mode of operation where a device configures its network interface to receive every frame on the LAN, regardless of its destination address.

Propagation delay
: The signal transit time through a cable, network segment, or device.

Protocol
: A set of agreed-upon rules and message formats for exchanging information among devices on a network.

QoS
: Quality of Service. QoS is typically achieved by providing different levels of service priority for packet transmission such that, in event of congestion on a switch port, higher-priority packets are served first and lower-priority packets are more likely to be dropped. Class of Service bits are used to

3. IEEE Std 802.3-2012, p. 2.

provide priority tagging on Ethernet frames.

Receive collision
: A collision detected on a coaxial media segment by a device that isn't actively transmitting data. A collision on coaxial cables is sensed by monitoring the average voltage on the cable so a device that is not actively transmitting can still detect a collision. When a receive collision is detected by an Ethernet repeater, it will transmit a collision enforcement jam signal on all other ports.

Repeater
: A physical layer device used to interconnect LAN segments based on the same LAN technology and using the same data rate. An Ethernet repeater can only link Ethernet segments that are all operating in half-duplex mode and at the same speed.

RJ
: Registered Jack. A term from the telephone industry, used for jacks (connectors) registered for use with particular types of telephone services.

RJ45
: An eight-pin modular connector used on twisted-pair links. Officially, an RJ45 connector is a telephone connector designed for voice-grade circuits. RJ45-style connectors with improved signal handling characteristics are called eight-pin connectors in the standards documents, but most people continue to use the RJ45 name for all eight-pin connectors.

Router
: A device or process based on Layer 3 network protocols used to interconnect networks at the network layer.

RSTP
: Rapid Spanning Tree Protocol. Initially defined in the 802.1w supplement to the 802.1D standard, RSTP is an improved version of the Spanning Tree Protocol (STP) that is interoperable with the classic STP. RSTP provides significantly faster spanning tree convergence in a Layer 2 network composed of Ethernet switches.

SC
: Subscriber Connector. This is a type of fiber optic connector used in 100BASE-FX and 1000BASE-LX/SX fiber optic media systems. The connector is designed to be pushed into place, automatically seating itself.

Segment
: An Ethernet media segment made up of a cable section for carrying Ethernet signals.

Signal crossover
: On a twisted-pair or fiber optic link segment, the transmit signals at one end of the segment must be connected to the receiver at the other end of the segment, and vice versa. This is referred to as signal crossover.

Silver satin
: The name for the silver-gray voice-grade patch cable used to connect a telephone to a wall jack. Typical silver satin patch cables do not have twisted-pair wires, which makes them unsuitable for use in an Ethernet system. The lack of twisted pairs will result in high levels of crosstalk, which can lead to slow performance on a 10BASE-T link and complete link failures on faster links.

Slot time
: A unit of time used in the Media Access Control (MAC) protocol for Ethernets.

SNMP
: Simple Network Management Protocol. A protocol specified by the Internet Engineering Task Force (IETF) for exchanging network management information between network devices and network management stations.

SQE
: Signal Quality Error. This signal indicates the detection of a collision on the medium by the transceiver. The original DIX Ethernet standard referred to this signal as Collision Presence; however, the name was changed to SQE in the IEEE 802.3 specifications.

SQE test
: This signal tests the SQE detection and signaling circuits. The original DIX Ethernet standard referred to this as the Collision Presence Test, also known as "heartbeat." The name was changed to SQE Test in the IEEE 802.3 specifications.

ST
: Straight Tip. Developed by AT&T, this is a type of fiber optic connector used in 10BASE-FL and FOIRL links. The male end of this connector has an inner sleeve with a slot cut into it, and an outer ring with a bayonet latch. The inner sleeve is aligned with a mating key in the socket, and the outer ring is turned to complete the bayonet latch.

Star topology
: A network topology in which each node on the network is connected directly to a central hub.

Station
: A unique, addressable device on a network.

STP
: Spanning Tree Protocol. A network protocol used on bridges to ensure a loop-free topology in a local area network.

Switch
: Another name for a bridge, which is a device that interconnects network segments at the data link layer of network operations (Layer 2). Switches provide multiple ports for connections to network devices.

Tap
: A method of connecting a transceiver to a thick coaxial cable by drilling a hole in the cable and installing a transceiver tap connection.

Telco connector
: See *50-pin connector*.

Throughput
: The rate at which usable data can be sent over the channel. While an Ethernet channel may operate at any of the Ethernet speeds, the throughput in terms of usable data will be less than the rated speed due to the number of bits required for framing and other channel overhead.

TIA/EIA
: Telecommunications Industry Association/Electronics Industry Association (TIA/EIA). An organization that specifies commercial building telecommunications cable standards, including the cable category specifications.

Topology
: The physical or logical layout of a network.

Station
: A unique, addressable device on a network.

Terminator
: A resistor used at the end of metallic baseband LAN cables to minimize reflections.

Transceiver
: A combination of the words *trans*-mitter and re-*ceiver*. A transceiver is the set of electronics that sends and receives signals on an Ethernet media system. Transceivers may be small outboard devices, or may be built into an Ethernet port.

Transceiver cable
: See *AUI cable*.

Twisted-pair cable
: A multiple-conductor cable whose component wires are paired together, twisted, and enclosed in a single jacket. A typical Category 5 twisted-pair segment is composed of a cable with four twisted pairs contained in a single jacket. Each pair consists of two insulated copper wires that are twisted together.

USOC
: Universal Service Order Code (pronounced "U-Sock"). An old Bell System term used to identify a particular service or device offered under tariff. Often used to refer to an old cable color code scheme that was current when USOC codes were in use.

Voice-grade
: A term for twisted-pair cable used in telephone systems to carry voice signals.

VLAN
: Virtual LAN. A method of grouping together one or more ports in a switch so that they function as a single "virtual" network. All ports within a given VLAN are members of the same broadcast domain.

Wiring closet
: Also called a telecommunications closet. A room that contains one or more wire distribution racks and panels used to connect various cables together (via patch cables) to form physical networks.

Index

Symbols

We'd like to hear your suggestions for improving our indexes. Send email to index@oreilly.com.

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

About the Authors

Charles E. Spurgeon is a senior technology architect at the University of Texas at Austin, where he works on a campus network system serving over 70,000 users in 200 buildings on two campuses. He has developed and managed large campus networks for many years, beginning at Stanford University, where he worked with a group that built the prototype Ethernet routers that became the founding technology for Cisco Systems. Charles, who attended Wesleyan University, lives in Austin, Texas, with his wife, Joann Zimmerman, and their cat, Mona.

Joann Zimmerman is a former software engineer with a doctorate in art history from the University of Texas at Austin. She has written and documented compilers, software tools, and network monitoring software, and has been a creator of the build and configuration management process for several companies. The author of papers in software engineering and Renaissance art history, she currently has multiple fantasy novels in process.

Colophon

The animal on the cover of *Ethernet: The Definitive Guide* is an octopus. The octopus is a member of the class *Cephalopoda*, which also includes squid, cuttlefish, and nautili. However, unlike other cephalopods, the octopus's shell is entirely absent. Species of octopus vary in size from under an inch (the Californian *Enteroctopus micropyrsus*) to 30 feet in length (the Giant Pacific *Octopus dofleini*). Like its cousin the squid, the octopus can release a noxious ink when disturbed. Octopodes vary in color from pink to brown, but are able to change their skin's complexion when threatened using special pigment cells called chromatophores.

Octopodes catch their prey—primarily crabs, lobsters, and other smaller sea creatures—with their suckered tentacles. Many species are aided by a poison these sucker cups secrete; the venom of one Australian species is so potent that it can be deadly to humans.

Octopodes are considered the most intelligent invertebrate species. They have both short- and long-term memory and have shown trial-and-error learning skills, retaining the problem-solving capabilities gained through experience. Their sucker cups are very sensitive; a sightless octopus can differentiate between various shapes and sizes of objects just as well as a sighted one.

The cover image is a 19th-century engraving from the Dover Pictorial Archive. The cover fonts are URW Typewriter and Guardian Sans. The text font is Adobe Minion Pro; the heading font is Adobe Myriad Condensed; and the code font is Dalton Maag's Ubuntu Mono.