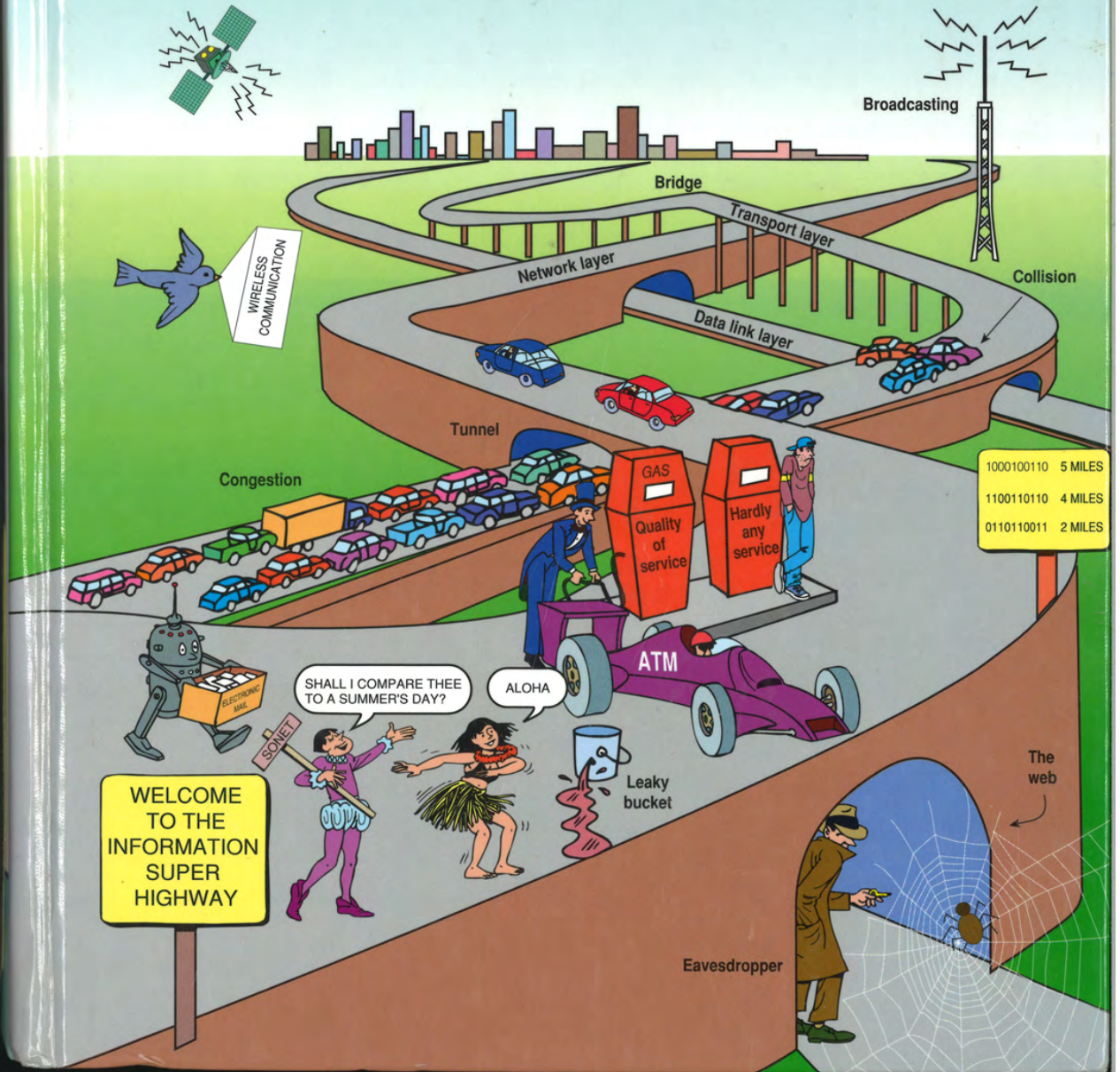


COMPUTER NETWORKS

ANDREW S. TANENBAUM



Library of Congress Cataloging in Publication Data

Tanenbaum, Andrew S. 1944-.

Computer networks / Andrew S. Tanenbaum. -- 3rd ed.

p. cm.

Includes bibliographical references and index.

ISBN 0-13-349945-6

1. Computer networks. I. Title.

TK5105.5.T36 1996

004.6--dc20

96-4121

CIP

Editorial/production manager: *Camille Trentacoste*

Interior design and composition: *Andrew S. Tanenbaum*

Cover design director: *Jerry Votta*

Cover designer: *Don Martinetti, DM Graphics, Inc.*

Cover concept: *Andrew S. Tanenbaum, from an idea by Marilyn Tremaine*

Interior graphics: *Hadel Studio*

Manufacturing manager: *Alexis R. Heydt*

Acquisitions editor: *Mary Franz*

Editorial Assistant: *Noreen Regina*



© 1996 by Prentice Hall PTR

Prentice-Hall, Inc.

A Simon & Schuster Company

Upper Saddle River, New Jersey 07458

The publisher offers discounts on this book when ordered in bulk quantities. For more information, contact:

Corporate Sales Department, Prentice Hall PTR, One Lake Street, Upper Saddle River, NJ 07458.

Phone: (800) 382-3419; Fax: (201) 236-7141. E-mail: corpsales@prehall.com

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

All product names mentioned herein are the trademarks of their respective owners.

Printed in the United States of America

10 9

ISBN 0-13-349945-6

Prentice-Hall International (UK) Limited, *London*

Prentice-Hall of Australia Pty. Limited, *Sydney*

Prentice-Hall Canada Inc., *Toronto*

Prentice-Hall Hispanoamericana, S.A., *Mexico*

Prentice-Hall of India Private Limited, *New Delhi*

Prentice-Hall of Japan, Inc., *Tokyo*

Simon & Schuster Asia Pte. Ltd., *Singapore*

Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

1.3.1. Protocol Hierarchies

To reduce their design complexity, most networks are organized as a series of **layers** or **levels**, each one built upon the one below it. The number of layers, the name of each layer, the contents of each layer, and the function of each layer differ from network to network. However, in all networks, the purpose of each layer is to offer certain services to the higher layers, shielding those layers from the details of how the offered services are actually implemented.

Layer n on one machine carries on a conversation with layer n on another machine. The rules and conventions used in this conversation are collectively known as the layer n **protocol**. Basically, a protocol is an agreement between the communicating parties on how communication is to proceed. As an analogy, when a woman is introduced to a man, she may choose to stick out her hand. He, in turn, may decide either to shake it or kiss it, depending, for example, on whether she is an American lawyer at a business meeting or a European princess at a formal ball. Violating the protocol will make communication more difficult, if not impossible.

A five-layer network is illustrated in Fig. 1-9. The entities comprising the corresponding layers on different machines are called **peers**. In other words, it is the peers that communicate using the protocol.

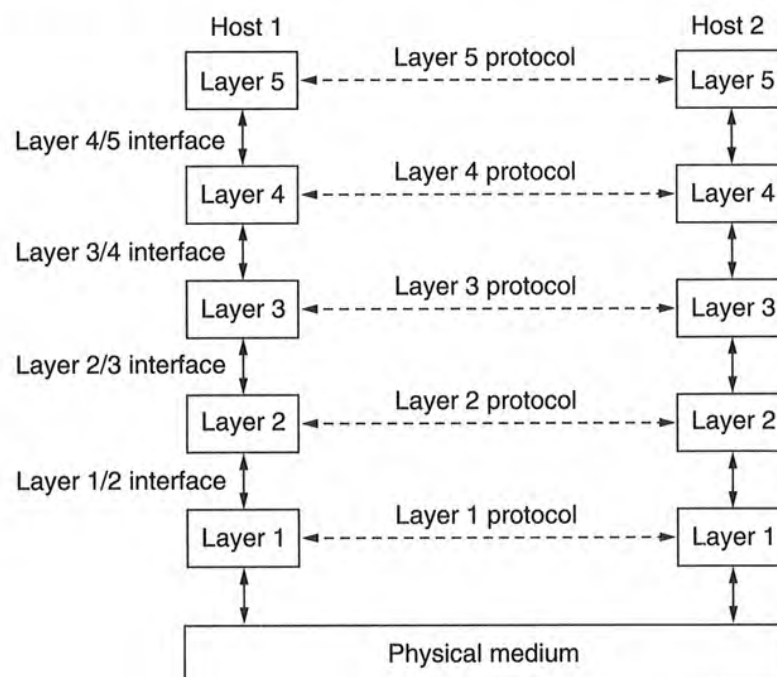


Fig. 1-9. Layers, protocols, and interfaces.

In reality, no data are directly transferred from layer n on one machine to layer n on another machine. Instead, each layer passes data and control

are free to change their protocols at will, provided they do not change the service visible to their users. In this way, the service and the protocol are completely decoupled.

An analogy with programming languages is worth making. A service is like an abstract data type or an object in an object-oriented language. It defines operations that can be performed on an object but does not specify how these operations are implemented. A protocol relates to the *implementation* of the service and as such is not visible to the user of the service.

Many older protocols did not distinguish the service from the protocol. In effect, a typical layer might have had a service primitive SEND PACKET with the user providing a pointer to a fully assembled packet. This arrangement meant that all changes to the protocol were immediately visible to the users. Most network designers now regard such a design as a serious blunder.

1.4. REFERENCE MODELS

Now that we have discussed layered networks in the abstract, it is time to look at some examples. In the next two sections we will discuss two important network architectures, the OSI reference model and the TCP/IP reference model.

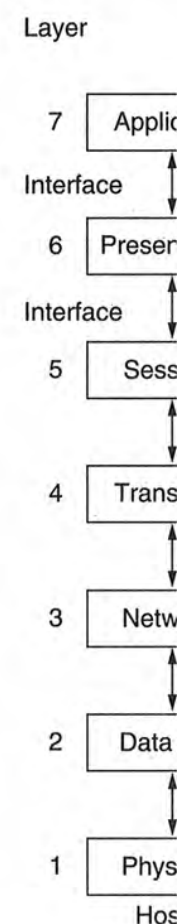
1.4.1. The OSI Reference Model

The OSI model is shown in Fig. 1-16 (minus the physical medium). This model is based on a proposal developed by the International Standards Organization (ISO) as a first step toward international standardization of the protocols used in the various layers (Day and Zimmermann, 1983). The model is called the **ISO OSI (Open Systems Interconnection) Reference Model** because it deals with connecting open systems—that is, systems that are open for communication with other systems. We will usually just call it the OSI model for short.

The OSI model has seven layers. The principles that were applied to arrive at the seven layers are as follows:

1. A layer should be created where a different level of abstraction is needed.
2. Each layer should perform a well defined function.
3. The function of each layer should be chosen with an eye toward defining internationally standardized protocols.
4. The layer boundaries should be chosen to minimize the information flow across the interfaces.
5. The number of layers should be large enough that distinct functions need not be thrown together in the same layer out of necessity, and small enough that the architecture does not become unwieldy.

Below w
layer. Note
not specify t
what each la
layers, altho
been publish



The Physical

The phy
cation chann
sends a 1 bi

Below we will discuss each layer of the model in turn, starting at the bottom layer. Note that the OSI model itself is not a network architecture because it does not specify the exact services and protocols to be used in each layer. It just tells what each layer should do. However, ISO has also produced standards for all the layers, although these are not part of the reference model itself. Each one has been published as a separate international standard.

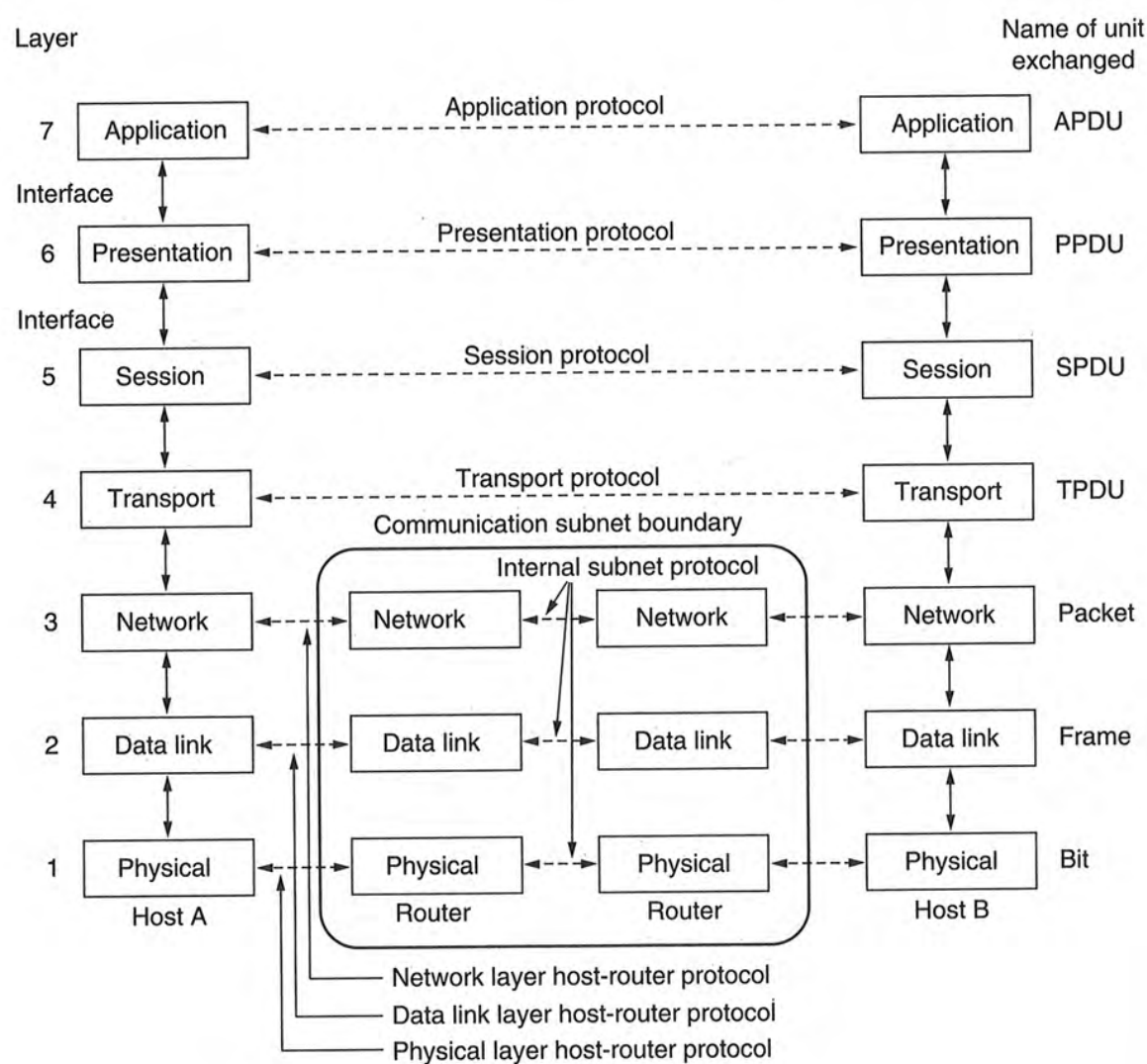


Fig. 1-16. The OSI reference model.

The Physical Layer

The **physical layer** is concerned with transmitting raw bits over a communication channel. The design issues have to do with making sure that when one side sends a 1 bit, it is received by the other side as a 1 bit, not as a 0 bit. Typical

questions here are how many volts should be used to represent a 1 and how many for a 0, how many microseconds a bit lasts, whether transmission may proceed simultaneously in both directions, how the initial connection is established and how it is torn down when both sides are finished, and how many pins the network connector has and what each pin is used for. The design issues here largely deal with mechanical, electrical, and procedural interfaces, and the physical transmission medium, which lies below the physical layer.

The Data Link Layer

The main task of the **data link layer** is to take a raw transmission facility and transform it into a line that appears free of undetected transmission errors to the network layer. It accomplishes this task by having the sender break the input data up into **data frames** (typically a few hundred or a few thousand bytes), transmit the frames sequentially, and process the **acknowledgement frames** sent back by the receiver. Since the physical layer merely accepts and transmits a stream of bits without any regard to meaning or structure, it is up to the data link layer to create and recognize frame boundaries. This can be accomplished by attaching special bit patterns to the beginning and end of the frame. If these bit patterns can accidentally occur in the data, special care must be taken to make sure these patterns are not incorrectly interpreted as frame delimiters.

A noise burst on the line can destroy a frame completely. In this case, the data link layer software on the source machine can retransmit the frame. However, multiple transmissions of the same frame introduce the possibility of duplicate frames. A duplicate frame could be sent if the acknowledgement frame from the receiver back to the sender were lost. It is up to this layer to solve the problems caused by damaged, lost, and duplicate frames. The data link layer may offer several different service classes to the network layer, each of a different quality and with a different price.

Another issue that arises in the data link layer (and most of the higher layers as well) is how to keep a fast transmitter from drowning a slow receiver in data. Some traffic regulation mechanism must be employed to let the transmitter know how much buffer space the receiver has at the moment. Frequently, this flow regulation and the error handling are integrated.

If the line can be used to transmit data in both directions, this introduces a new complication that the data link layer software must deal with. The problem is that the acknowledgement frames for *A* to *B* traffic compete for the use of the line with data frames for the *B* to *A* traffic. A clever solution (piggybacking) has been devised; we will discuss it in detail later.

Broadcast networks have an additional issue in the data link layer: how to control access to the shared channel. A special sublayer of the data link layer, the medium access sublayer, deals with this problem.

The Network

The net
A key design
tion. Routes
rarely chang
example a t
mined anew

If too ma
each other's
belongs to th

Since th
efforts, there
the very leas
sent by each
national bor
complicated.

When a
tion, many p
be different
because it is
work layer to
interconnect

In broad
often thin or

The Transp

The basi
layer, split it
ensure that t
must be done
itable change

Under no
nection for e
port connecti
create multip
nections to i
network con
transport con
cases, the tra
session layer.

The trans

The Network Layer

The **network layer** is concerned with controlling the operation of the subnet. A key design issue is determining how packets are routed from source to destination. Routes can be based on static tables that are "wired into" the network and rarely changed. They can also be determined at the start of each conversation, for example a terminal session. Finally, they can be highly dynamic, being determined anew for each packet, to reflect the current network load.

If too many packets are present in the subnet at the same time, they will get in each other's way, forming bottlenecks. The control of such congestion also belongs to the network layer.

Since the operators of the subnet may well expect remuneration for their efforts, there is often some accounting function built into the network layer. At the very least, the software must count how many packets or characters or bits are sent by each customer, to produce billing information. When a packet crosses a national border, with different rates on each side, the accounting can become complicated.

When a packet has to travel from one network to another to get to its destination, many problems can arise. The addressing used by the second network may be different from the first one. The second one may not accept the packet at all because it is too large. The protocols may differ, and so on. It is up to the network layer to overcome all these problems to allow heterogeneous networks to be interconnected.

In broadcast networks, the routing problem is simple, so the network layer is often thin or even nonexistent.

The Transport Layer

The basic function of the **transport layer** is to accept data from the session layer, split it up into smaller units if need be, pass these to the network layer, and ensure that the pieces all arrive correctly at the other end. Furthermore, all this must be done efficiently, and in a way that isolates the upper layers from the inevitable changes in the hardware technology.

Under normal conditions, the transport layer creates a distinct network connection for each transport connection required by the session layer. If the transport connection requires a high throughput, however, the transport layer might create multiple network connections, dividing the data among the network connections to improve throughput. On the other hand, if creating or maintaining a network connection is expensive, the transport layer might multiplex several transport connections onto the same network connection to reduce the cost. In all cases, the transport layer is required to make the multiplexing transparent to the session layer.

The transport layer also determines what type of service to provide the session

layer, and ultimately, the users of the network. The most popular type of transport connection is an error-free point-to-point channel that delivers messages or bytes in the order in which they were sent. However, other possible kinds of transport service are transport of isolated messages with no guarantee about the order of delivery, and broadcasting of messages to multiple destinations. The type of service is determined when the connection is established.

The transport layer is a true end-to-end layer, from source to destination. In other words, a program on the source machine carries on a conversation with a similar program on the destination machine, using the message headers and control messages. In the lower layers, the protocols are between each machine and its immediate neighbors, and not by the ultimate source and destination machines, which may be separated by many routers. The difference between layers 1 through 3, which are chained, and layers 4 through 7, which are end-to-end, is illustrated in Fig. 1-16.

Many hosts are multiprogrammed, which implies that multiple connections will be entering and leaving each host. There needs to be some way to tell which message belongs to which connection. The transport header (H_4 in Fig. 1-11) is one place this information can be put.

In addition to multiplexing several message streams onto one channel, the transport layer must take care of establishing and deleting connections across the network. This requires some kind of naming mechanism, so that a process on one machine has a way of describing with whom it wishes to converse. There must also be a mechanism to regulate the flow of information, so that a fast host cannot overrun a slow one. Such a mechanism is called **flow control** and plays a key role in the transport layer (also in other layers). Flow control between hosts is distinct from flow control between routers, although we will later see that similar principles apply to both.

The Session Layer

The session layer allows users on different machines to establish **sessions** between them. A session allows ordinary data transport, as does the transport layer, but it also provides enhanced services useful in some applications. A session might be used to allow a user to log into a remote timesharing system or to transfer a file between two machines.

One of the services of the session layer is to manage dialogue control. Sessions can allow traffic to go in both directions at the same time, or in only one direction at a time. If traffic can only go one way at a time (analogous to a single railroad track), the session layer can help keep track of whose turn it is.

A related session service is **token management**. For some protocols, it is essential that both sides do not attempt the same operation at the same time. To manage these activities, the session layer provides tokens that can be exchanged. Only the side holding the token may perform the critical operation.

Another might occur 1-hour message transfer with time as we check point after the last

The Presentation Layer

The presentation layer is concerned with the efficient exchange of data between each user and the system. It is just interested in the data concerned.

A typical example is agreed upon. They exchange invoices. The numbers, and computers have Unicode), in order to manage the data with a standard manages the data inside the system.

The Application Layer

The application layer is needed. For the world. Communication network with escape sequence.

One way is a terminal type that external type work virtual software cursor then. Another

of transport
ages or bytes
of transport
the order of
type of ser-

stination. In
sation with a
ers and con-
chine and its
on machines,
een layers 1
nd-to-end, is

connections
to tell which
Fig. 1-11) is

channel, the
ns across the
ocess on one

There must
t host cannot
l plays a key
ween hosts is
e that similar

lish sessions
the transport
tions. A ses-
system or to

control. Ses-
r in only one
us to a single
is.

otocols, it is
me time. To
e exchanged.

Another session service is **synchronization**. Consider the problems that might occur when trying to do a 2-hour file transfer between two machines with a 1-hour mean time between crashes. After each transfer was aborted, the whole transfer would have to start over again and would probably fail again the next time as well. To eliminate this problem, the session layer provides a way to insert checkpoints into the data stream, so that after a crash, only the data transferred after the last checkpoint have to be repeated.

The Presentation Layer

The **presentation layer** performs certain functions that are requested sufficiently often to warrant finding a general solution for them, rather than letting each user solve the problems. In particular, unlike all the lower layers, which are just interested in moving bits reliably from here to there, the presentation layer is concerned with the syntax and semantics of the information transmitted.

A typical example of a presentation service is encoding data in a standard agreed upon way. Most user programs do not exchange random binary bit strings. They exchange things such as people's names, dates, amounts of money, and invoices. These items are represented as character strings, integers, floating-point numbers, and data structures composed of several simpler items. Different computers have different codes for representing character strings (e.g., ASCII and Unicode), integers (e.g., one's complement and two's complement), and so on. In order to make it possible for computers with different representations to communicate, the data structures to be exchanged can be defined in an abstract way, along with a standard encoding to be used "on the wire." The presentation layer manages these abstract data structures and converts from the representation used inside the computer to the network standard representation and back.

The Application Layer

The **application layer** contains a variety of protocols that are commonly needed. For example, there are hundreds of incompatible terminal types in the world. Consider the plight of a full screen editor that is supposed to work over a network with many different terminal types, each with different screen layouts, escape sequences for inserting and deleting text, moving the cursor, etc.

One way to solve this problem is to define an abstract **network virtual terminal** that editors and other programs can be written to deal with. To handle each terminal type, a piece of software must be written to map the functions of the network virtual terminal onto the real terminal. For example, when the editor moves the virtual terminal's cursor to the upper left-hand corner of the screen, this software must issue the proper command sequence to the real terminal to get its cursor there too. All the virtual terminal software is in the application layer.

Another application layer function is file transfer. Different file systems have

different file naming conventions, different ways of representing text lines, and so on. Transferring a file between two different systems requires handling these and other incompatibilities. This work, too, belongs to the application layer, as do electronic mail, remote job entry, directory lookup, and various other general-purpose and special-purpose facilities.

Data Transmission in the OSI Model

Figure 1-17 shows an example of how data can be transmitted using the OSI model. The sending process has some data it wants to send to the receiving process. It gives the data to the application layer, which then attaches the application header, *AH* (which may be null), to the front of it and gives the resulting item to the presentation layer.

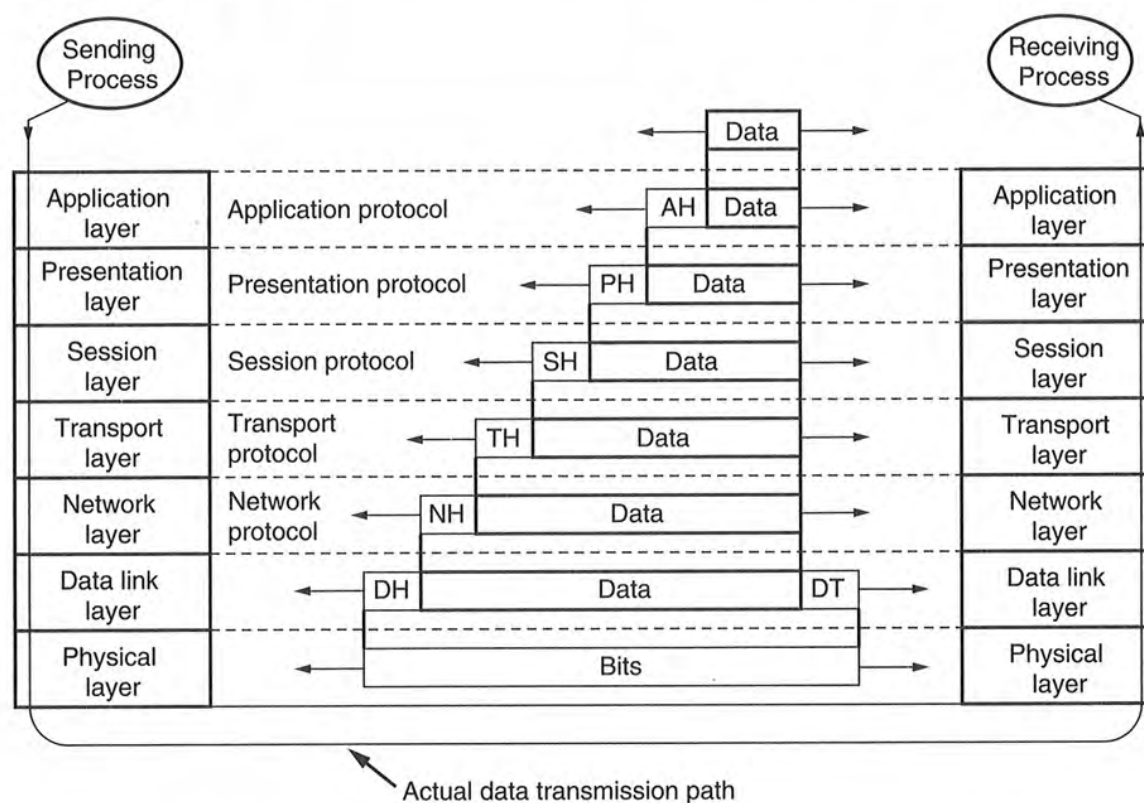


Fig. 1-17. An example of how the OSI model is used. Some of the headers may be null. (Source: H.C. Folts. Used with permission.)

The presentation layer may transform this item in various ways and possibly add a header to the front, giving the result to the session layer. It is important to realize that the presentation layer is not aware of which portion of the data given to it by the application layer is *AH*, if any, and which is true user data.

This process is repeated until the data reach the physical layer, where they are actually transmitted to the receiving machine. On that machine the various

headers are it finally arr

The key Fig. 1-17, e ing transpor a transport l view, the fa own machin speaking di addressing t to his transla

1.4.2. The

Let us n the grandpa worldwide I it is useful research net ally connect telephone li protocols ha needed. Th was one of later becam tocols. It w in (Leiner e (Clark, 1988

Given tl work gatew goal was the conversation remain intact even if som out of opera tions with d to real-time

The Intern

All thes on a connec the linchpin

An implicit assumption in the above discussion is that the power levels of all stations are the same as perceived by the receiver. CDMA is typically used for wireless systems with a fixed base station and many mobile stations at varying distances from it. The power levels received at the base station depend on how far away the transmitters are. A good heuristic here is for each mobile station to transmit to the base station at the inverse of the power level it receives from the base station, so a mobile station receiving a weak signal from the base will use more power than one getting a strong signal. The base station can also give explicit commands to the mobile stations to increase or decrease their transmission power.

We have also assumed that the receiver knows who the sender is. In principle, given enough computing capacity, the receiver can listen to all the senders at once by running the decoding algorithm for each of them in parallel. In real life, suffice it to say that this is easier said than done. CDMA also has many other complicating factors that have been glossed over in this brief introduction. Nevertheless, CDMA is a clever scheme that is being rapidly introduced for wireless mobile communication.

Readers with a solid electrical engineering background who want to gain a deeper understanding of CDMA should read (Viterbi, 1995). An alternative spreading scheme, in which the spreading is over time rather than frequency, is described in (Crespo et al., 1995).

4.3. IEEE STANDARD 802 FOR LANS AND MANS

We have now finished our general discussion of abstract channel allocation protocols, so it is time to see how these principles apply to real systems, in particular, LANs. As discussed in Sec. 1.7.2, IEEE has produced several standards for LANs. These standards, collectively known as **IEEE 802**, include CSMA/CD, token bus, and token ring. The various standards differ at the physical layer and MAC sublayer but are compatible at the data link layer. The IEEE 802 standards have been adopted by ANSI as American National Standards, by NIST as government standards, and by ISO as international standards (known as ISO 8802). They are surprisingly readable (as standards go).

The standards are divided into parts, each published as a separate book. The 802.1 standard gives an introduction to the set of standards and defines the interface primitives. The 802.2 standard describes the upper part of the data link layer, which uses the **LLC (Logical Link Control)** protocol. Parts 802.3 through 802.5 describe the three LAN standards, the CSMA/CD, token bus, and token ring standards, respectively. Each standard covers the physical layer and MAC sublayer protocol. The next three sections cover these three systems. Additional information can be found in (Stallings, 1993b).

including collisions and binary backoff. Successful frames make it to the switch, and are treated there like any other incoming frames: they are switched to the correct output line over the high-speed backplane. If all the input ports are connected to hubs, rather than to individual stations, the switch just becomes an 802.3 to 802.3 bridge. We will study bridges later in this chapter.

4.3.2. IEEE Standard 802.4: Token Bus

Although 802.3 is widely used in offices, during the development of the 802 standard, people from General Motors and other companies interested in factory automation had serious reservations about it. For one thing, due to the probabilistic MAC protocol, with a little bad luck a station might have to wait arbitrarily long to send a frame (i.e., the worst case is unbounded). For another, 802.3 frames do not have priorities, making them unsuited for real-time systems in which important frames should not be held up waiting for unimportant frames.

A simple system with a known worst case is a ring in which the stations take turns sending frames. If there are n stations and it takes T sec to send a frame, no frame will ever have to wait more than nT sec to be sent. The factory automation people in the 802 committee liked the conceptual idea of a ring but did not like the physical implementation because a break in the ring cable would bring the whole network down. Furthermore, they noted that a ring is a poor fit to the linear topology of most assembly lines. As a result, a new standard was developed, having the robustness of the 802.3 broadcast cable, but the known worst-case behavior of a ring.

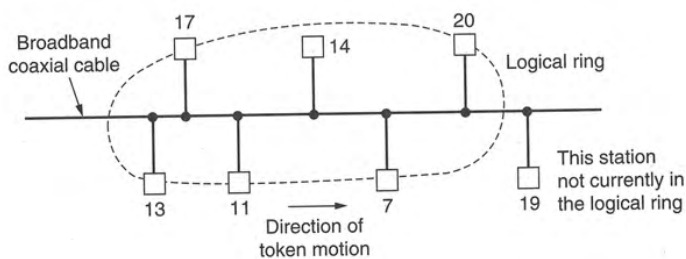


Fig. 4-25. A token bus.

This standard, 802.4 (Dirvin and Miller, 1986; and IEEE, 1985b), describes a LAN called a **token bus**. Physically, the token bus is a linear or tree-shaped cable onto which the stations are attached. Logically, the stations are organized into a ring (see Fig. 4-25), with each station knowing the address of the station to its "left" and "right." When the logical ring is initialized, the highest numbered station may send the first frame. After it is done, it passes permission to its immediate neighbor by sending the neighbor a special control frame called a **token**. The

token propagates around the logical ring, with only the token holder being permitted to transmit frames. Since only one station at a time holds the token, collisions do not occur.

An important point to realize is that the physical order in which the stations are connected to the cable is not important. Since the cable is inherently a broadcast medium, each station receives each frame, discarding those not addressed to it. When a station passes the token, it sends a token frame specifically addressed to its logical neighbor in the ring, irrespective of where that station is physically located on the cable. It is also worth noting that when stations are first powered on, they will not be in the ring (e.g., stations 14 and 19 in Fig. 4-25), so the MAC protocol has provisions for adding stations to, and deleting stations from, the ring.

The 802.4 MAC protocol is very complex, with each station having to maintain ten different timers and more than two dozen internal state variables. The 802.4 standard is much longer than 802.3, filling more than 200 pages. The two standards are also quite different in style, with 802.3 giving the protocols as Pascal procedures, whereas 802.4 gives them as finite state machines, with the actions written in Ada[®].

For the physical layer, the token bus uses the 75-ohm broadband coaxial cable used for cable television. Both single- and dual-cable systems are allowed, with or without head-ends. Three different analog modulation schemes are permitted: phase continuous frequency shift keying, phase coherent frequency shift keying, and multilevel duobinary amplitude modulated phase shift keying. Speeds of 1, 5, and 10 Mbps are possible. Furthermore, the modulation schemes not only provide ways to represent 0, 1, and idle on the cable, but also three other symbols used for network control. All in all, the physical layer is totally incompatible with 802.3, and a lot more complicated.

The Token Bus MAC Sublayer Protocol

When the ring is initialized, stations are inserted into it in order of station address, from highest to lowest. Token passing is also done from high to low addresses. Each time a station acquires the token, it can transmit frames for a certain amount of time; then it must pass the token on. If the frames are short enough, several consecutive frames may be sent. If a station has no data, it passes the token immediately upon receiving it.

The token bus defines four priority classes, 0, 2, 4, and 6 for traffic, with 0 the lowest and 6 the highest. It is easiest to think of each station internally being divided into four substations, one at each priority level. As input comes in to the MAC sublayer from above, the data are checked for priority and routed to one of the four substations. Thus each substation maintains its own queue of frames to be transmitted.

When the token comes into the station over the cable, it is passed internally to the priority 6 substation, which may begin transmitting frames, if it has any.

When it is done (or when its timer expires), the token is passed internally to the priority 4 substation, which may then transmit frames until its timer expires, at which point the token is passed internally to the priority 2 substation. This process is repeated until either the priority 0 substation has sent all its frames or its timer has expired. Either way, at this point the token is sent to the next station in the ring.

Without getting into all the details of how the various timers are managed, it should be clear that by setting the timers properly, we can ensure that a guaranteed fraction of the total token-holding time can be allocated to priority 6 traffic. The lower priorities will have to live with what is left over. If the higher priority substations do not need all of their allocated time, the lower priority substations can have the unused portion, so it is not wasted.

This priority scheme, which guarantees priority 6 traffic a known fraction of the network bandwidth, can be used to implement real-time traffic. For example, suppose the parameters of a 50-station network running at 10 Mbps have been adjusted to give priority 6 traffic 1/3 of the bandwidth. Then each station has a guaranteed 67 kbps for priority 6 traffic. This bandwidth could be used to synchronize robots on an assembly line or carry one digital voice channel per station, with a little left over for control information.

The token bus frame format is shown in Fig. 4-26. It is unfortunately different from the 802.3 frame format. The preamble is used to synchronize the receiver's clock, as in 802.3, except that here it may be as short as 1 byte. The *Starting delimiter* and *Ending delimiter* fields are used to mark the frame boundaries. Both of these fields contain analog encoding of symbols other than 0s and 1s, so that they cannot occur accidentally in the user data. As a result, no length field is needed.

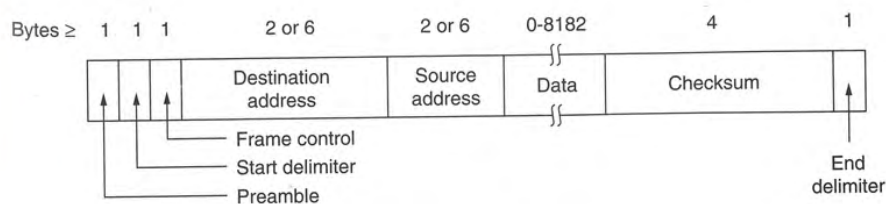


Fig. 4-26. The 802.4 frame format.

The *Frame control* field is used to distinguish data frames from control frames. For data frames, it carries the frame's priority. It can also carry an indicator requiring the destination station to acknowledge correct or incorrect receipt of the frame. Without this indicator, the destination would not be allowed to send anything because it does not have the token. This indicator turns the token bus into something resembling the acknowledgement scheme of Tokoro and Tamaru.

For control frames, the *Frame control* field is used to specify the frame type.

The allowed types include token passing and various ring maintenance frames, including the mechanism for letting new stations enter the ring, the mechanism for allowing stations to leave the ring, and so on. Note that the 802.3 protocol does not have any control frames. All the MAC layer does there is provide a way to get frames onto the cable; it does not care what is in them.

The *Destination address* and *Source address* fields are the same as in 802.3 (yes, the two groups did talk to each other; no, they did not agree on very much). As in 802.3, a given network must use all 2-byte addresses or all 6-byte addresses, not a mixture on the same cable. The initial 802.4 standard allows either size. The individual and group addressing and the local and global address assignments are identical to 802.3.

The *Data* field may be up to 8182 bytes long when 2-byte addresses are used, and up to 8174 bytes long when 6-byte addresses are used. This is more than five times as long as the maximum 802.3 frame, which was made short to prevent one station from hogging the channel too long. With the token bus, the timers can be used as an antihogging measure, but it is nice to be able to send long frames when real-time traffic is not an issue. The *Checksum* is used to detect transmission errors. It uses the same algorithm and polynomial as 802.3.

The token bus control frames are shown in Fig. 4-27. They will be discussed below. The only one we have seen so far is the *token* frame, used to pass the token from station to station. Most of the rest relate to adding and deleting stations from the logical ring.

Frame control field	Name	Meaning
00000000	Claim_token	Claim token during ring initialization
00000001	Solicit_successor_1	Allow stations to enter the ring
00000010	Solicit_successor_2	Allow stations to enter the ring
00000011	Who_follows	Recover from lost token
00000100	Resolve_contention	Used when multiple stations want to enter
00001000	Token	Pass the token
00001100	Set_successor	Allow station to leave the ring

Fig. 4-27. The token bus control frames.

Logical Ring Maintenance

From time to time, stations are powered on and want to join the ring. Others are turned off and want to leave. The MAC sublayer protocol provides a detailed specification of exactly how this is done while maintaining the known worst case bound on token rotation. Below we will just briefly sketch the mechanisms used.

Once the ring has been established, each station's interface maintains the addresses of the predecessor and successor stations internally. Periodically, the token holder sends one of the SOLICIT_SUCCESSOR frames shown in Fig. 4-27 to solicit bids from stations that wish to join the ring. The frame gives the sender's address and the successor's address. Stations inside that range may bid to enter (to keep the ring sorted in descending order of station address).

If no station bids to enter within a slot time (2τ , as in 802.3), the **response window** is closed and the token holder continues with its normal business. If exactly one station bids to enter, it is inserted into the ring and becomes the token holder's successor.

If two or more stations bid to enter, their frames will collide and be garbled, as in 802.3. The token holder then runs an arbitration algorithm, starting with the broadcast of a RESOLVE_CONTENTION frame. The algorithm is a variation of binary countdown, using two bits at a time.

Furthermore, all station interfaces maintain two random bits inside. These bits are used to delay all bids by 0, 1, 2, or 3 slot times, to further reduce contention. In other words, two stations only collide on a bid if the current two address bits being used are the same and they happen to have the same two random bits. To prevent stations that must wait 3 slot times from being at a permanent disadvantage, the random bits are regenerated every time they are used or periodically every 50 msec.

The solicitation of new stations may not interfere with the guaranteed worst case for token rotation. Each station has a timer that is reset whenever it acquires the token. When the token comes in, the old value of this timer (i.e., the previous token rotation time) is inspected just before the timer is reset. If it exceeds a certain threshold value, there has been too much traffic recently, so no bids may be solicited this time around. In any event, only one station may enter at each solicitation, to put a bound on how much time can be consumed in ring maintenance. No guarantee is provided for how long a station may have to wait to join the ring when traffic is heavy, but in practice it should not be more than a few seconds. This uncertainty is unfortunate, making 802.4 less suitable for real-time systems than its supporters often claim.

Leaving the ring is easy. A station, X , with successor S , and predecessor P , leaves the ring, by sending P a SET_SUCCESSOR frame telling it that henceforth its successor is S instead of X . Then X just stops transmitting.

Ring initialization is a special case of adding new stations. Consider an idle system with all stations powered off. When the first station comes on-line, it notices that there is no traffic for a certain period. Then it sends a CLAIM_TOKEN frame. Not hearing any competitors contending for the token, it creates a token and sets up a ring containing only itself. Periodically, it solicits bids for new stations to join. As new stations are powered on, they will respond to these bids and join the ring using the contention algorithm described above. Eventually, every station that wants to join the ring will be able to do so. If the first two stations are

powered on simultaneously, the protocol deals with this by letting them bid for the token using the standard modified binary countdown algorithm and the two random bits.

Due to transmission errors or hardware failures, problems can arise with the logical ring or the token. For example, if a station tries to pass the token to a station that has gone down, what happens? The solution is straightforward. After passing the token, a station listens to see if its successor either transmits a frame or passes the token. If it does neither, the token is passed a second time.

If that also fails, the station transmits a `WHO_FOLLOWS` frame specifying the address of its successor. When the failed station's successor sees a `WHO_FOLLOWS` frame naming its predecessor, it responds by sending a `SET_SUCCESSOR` frame to the station whose successor failed, naming itself as the new successor. In this way, the failed station is removed from the ring.

Now suppose that a station fails to pass the token to its successor and also fails to locate the successor's successor, which may also be down. It adopts a new strategy by sending a `SOLICIT_SUCCESSOR_2` frame to see if *anyone* else is still alive. Once again the standard contention protocol is run, with all stations that want to be in the ring now bidding for a place. Eventually, the ring is re-established.

Another kind of problem occurs if the token holder goes down and takes the token with it. This problem is solved using the ring initialization algorithm. Each station has a timer that is reset whenever a frame appears on the network. When this timer hits a threshold value, the station issues a `CLAIM_TOKEN` frame, and the modified binary countdown algorithm with random bits determines who gets the token.

Still another problem is multiple tokens. If a station holding the token notices a transmission from another station, it discards its token. If there were two, there would now be one. If there were more than two, this process would be repeated sooner or later until all but one were discarded. If, by accident, all the tokens are discarded, then the lack of activity will cause one or more stations to try to claim the token.

4.3.3. IEEE Standard 802.5: Token Ring

Ring networks have been around for many years (Pierce, 1972) and have long been used for both local and wide area networks. Among their many attractive features is the fact that a ring is not really a broadcast medium, but a collection of individual point-to-point links that happen to form a circle. Point-to-point links involve a well-understood and field-proven technology and can run on twisted pair, coaxial cable, or fiber optics. Ring engineering is also almost entirely digital, whereas 802.3, for example, has a substantial analog component for collision detection. A ring is also fair and has a known upper bound on channel access.

For these reasons, IBM chose the ring as its LAN and IEEE has included the **token ring** standard as 802.5 (IEEE, 1985c; Latif et al., 1992).

A major issue in the design and analysis of any ring network is the "physical length" of a bit. If the data rate of the ring is R Mbps, a bit is emitted every $1/R$ μ sec. With a typical signal propagation speed of about 200 m/ μ sec, each bit occupies $200/R$ meters on the ring. This means, for example, that a 1-Mbps ring whose circumference is 1000 meters can contain only 5 bits on it at once. The implications of the number of bits on the ring will become clearer later.

As mentioned above, a ring really consists of a collection of ring interfaces connected by point-to-point lines. Each bit arriving at an interface is copied into a 1-bit buffer and then copied out onto the ring again. While in the buffer, the bit can be inspected and possibly modified before being written out. This copying step introduces a 1-bit delay at each interface. A ring and its interfaces are shown in Fig. 4-28.

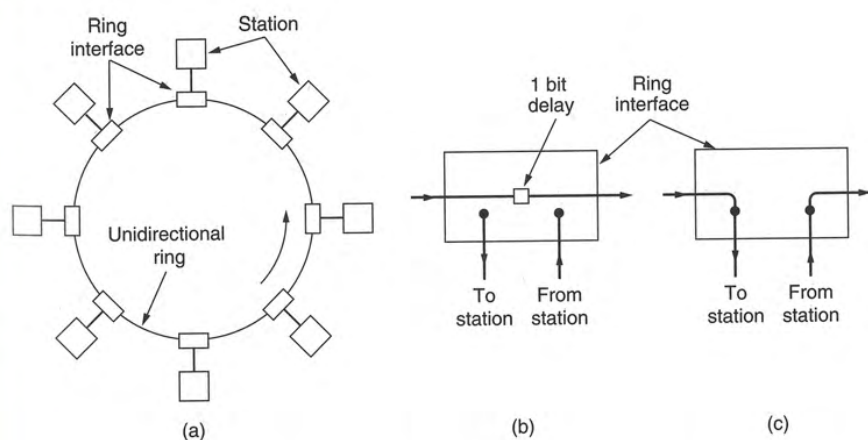


Fig. 4-28. (a) A ring network. (b) Listen mode. (c) Transmit mode.

In a token ring a special bit pattern, called the **token**, circulates around the ring whenever all stations are idle. When a station wants to transmit a frame, it is required to seize the token and remove it from the ring before transmitting. This action is done by inverting a single bit in the 3-byte token, which instantly changes it into the first 3 bytes of a normal data frame. Because there is only one token, only one station can transmit at a given instant, thus solving the channel access problem the same way the token bus solves it.

An implication of the token ring design is that the ring itself must have a sufficient delay to contain a complete token to circulate when all stations are idle. The delay has two components: the 1-bit delay introduced by each station, and the signal propagation delay. In almost all rings, the designers must assume that

stations may be powered down at various times, especially at night. If the interfaces are powered from the ring, shutting down the station has no effect on the interface, but if the interfaces are powered externally, they must be designed to connect the input to the output when power goes down, thus removing the 1-bit delay. The point here is that on a short ring an artificial delay may have to be inserted into the ring at night to ensure that a token can be contained on it.

Ring interfaces have two operating modes, listen and transmit. In listen mode, the input bits are simply copied to output, with a delay of 1 bit time, as shown in Fig. 4-28(b). In transmit mode, which is entered only after the token has been seized, the interface breaks the connection between input and output, entering its own data onto the ring. To be able to switch from listen to transmit mode in 1 bit time, the interface usually needs to buffer one or more frames itself rather than having to fetch them from the station on such short notice.

As bits that have propagated around the ring come back, they are removed from the ring by the sender. The sending station can either save them, to compare with the original data to monitor ring reliability, or discard them. Because the entire frame never appears on the ring at one instant, this ring architecture puts no limit on the size of the frames. After a station has finished transmitting the last bit of its last frame, it must regenerate the token. When the last bit of the frame has gone around and come back, it must be removed, and the interface must switch back into listen mode immediately, to avoid removing the token that might follow if no other station has removed it.

It is straightforward to handle acknowledgements on a token ring. The frame format need only include a 1-bit field for acknowledgements, initially zero. When the destination station has received a frame, it sets the bit. Of course, if the acknowledgement means that the checksum has been verified, the bit must follow the checksum, and the ring interface must be able to verify the checksum as soon as its last bit has arrived. When a frame is broadcast to multiple stations, a more complicated acknowledgement mechanism must be used (if any is used at all).

When traffic is light, the token will spend most of its time idly circulating around the ring. Occasionally a station will seize it, transmit a frame, and then output a new token. However, when the traffic is heavy, so that there is a queue at each station, as soon as a station finishes its transmission and regenerates the token, the next station downstream will see and remove the token. In this manner the permission to send rotates smoothly around the ring, in round-robin fashion. The network efficiency can begin to approach 100 percent under conditions of heavy load.

Now let us turn from token rings in general to the 802.5 standard in particular. At the physical layer, 802.5 calls for shielded twisted pairs running at 1 or 4 Mbps, although IBM later introduced a 16-Mbps version. Signals are encoded using differential Manchester encoding [see Fig. 4-20(c)] with high and low being positive and negative signals of absolute magnitude 3.0 to 4.5 volts. Normally, differential Manchester encoding uses high-low or low-high for each bit, but

802.5 also uses high-high and low-low in certain control bytes (e.g., to mark the start and end of a frame). These nondata signals always occur in consecutive pairs so as not to introduce a DC component into the ring voltage.

One problem with a ring network is that if the cable breaks somewhere, the ring dies. This problem can be solved very elegantly by the use of a **wire center**, as shown in Fig. 4-29. While logically still a ring, physically each station is connected to the wire center by a cable containing (at least) two twisted pairs, one for data to the station and one for data from the station.

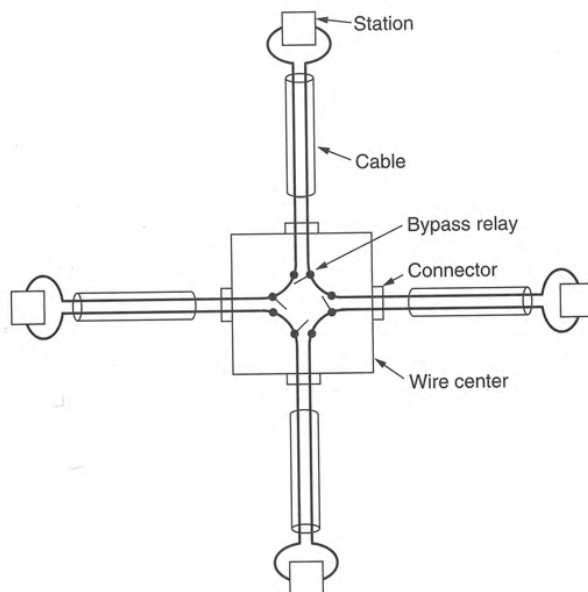


Fig. 4-29. Four stations connected via a wire center.

Inside the wire center are bypass relays that are energized by current from the stations. If the ring breaks or a station goes down, loss of the drive current will release the relay and bypass the station. The relays can also be operated by software to permit diagnostic programs to remove stations one at a time to find faulty stations and ring segments. The ring can then continue operation with the bad segment bypassed. Although the 802.5 standard does not formally require this kind of ring, often called a **star-shaped ring** (Saltzer et al., 1983), most 802.5 LANs, in fact, do use wire centers to improve their reliability and maintainability.

When a network consists of many clusters of stations far apart, a topology with multiple wire centers can be used. Just imagine that the cable to one of the stations in Fig. 4-29 were replaced by a cable to a distant wire center. Although logically all the stations are on the same ring, the wiring requirements are greatly

reduced. An 802.5 ring using a wire center has a similar topology to an 802.3 10Base-T hub-based network, but the formats and protocols are different.

The Token Ring MAC Sublayer Protocol

The basic operation of the MAC protocol is straightforward. When there is no traffic on the ring, a 3-byte token circulates endlessly, waiting for a station to seize it by setting a specific 0 bit to a 1 bit, thus converting the token into the start-of-frame sequence. The station then outputs the rest of a normal data frame, as shown in Fig. 4-30.

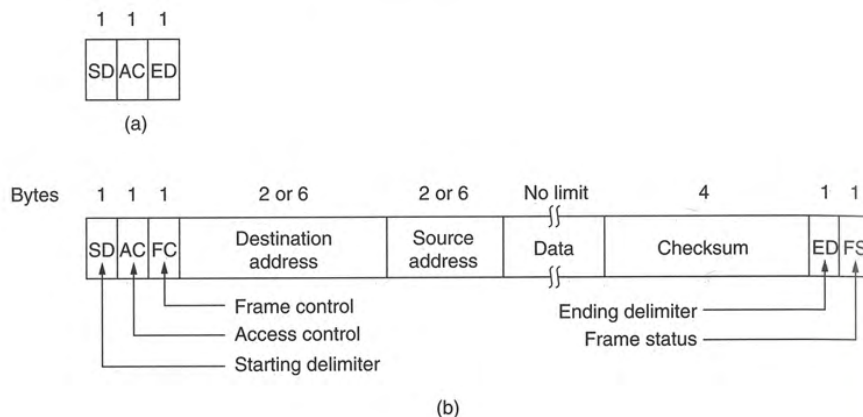


Fig. 4-30. (a) Token format. (b) Data frame format.

Under normal conditions, the first bit of the frame will go around the ring and return to the sender before the full frame has been transmitted. Only a very long ring will be able to hold even a short frame. Consequently, the transmitting station must drain the ring while it continues to transmit. As shown in Fig. 4-28(c), this means that the bits that have completed the trip around the ring come back to the sender and are removed.

A station may hold the token for the **token-holding time**, which is 10 msec unless an installation sets a different value. If there is enough time left after the first frame has been transmitted to send more frames, these may be sent as well. After all pending frames have been transmitted or the transmission of another frame would exceed the token-holding time, the station regenerates the 3-byte token frame and puts it out onto the ring.

The *Starting delimiter* and *Ending delimiter* fields of Fig. 4-30(b) mark the beginning and ending of the frame. Each contains invalid differential Manchester patterns (HH and LL) to distinguish them from data bytes. The *Access control*

byte contains the token bit, and also the *Monitor bit*, *Priority bits*, and *Reservation bits* (described below). The *Frame control* byte distinguishes data frames from various possible control frames.

Next come the *Destination address* and *Source address* fields, which are the same as in 802.3 and 802.4. These are followed by the data, which may be as long as necessary, provided that the frame can still be transmitted within the token-holding time. The *Checksum* field, like the destination and source addresses, is also the same as 802.3 and 802.4.

An interesting byte not present in the other two protocols is the *Frame status* byte. It contains the *A* and *C* bits. When a frame arrives at the interface of a station with the destination address, the interface turns on the *A* bit as it passes through. If the interface copies the frame to the station, it also turns on the *C* bit. A station might fail to copy a frame due to lack of buffer space or other reasons.

When the sending station drains the frame from the ring, it examines the *A* and *C* bits. Three combinations are possible:

1. $A = 0$ and $C = 0$: destination not present or not powered up.
2. $A = 1$ and $C = 0$: destination present but frame not accepted.
3. $A = 1$ and $C = 1$: destination present and frame copied.

This arrangement provides an automatic acknowledgement for each frame. If a frame is rejected but the station is present, the sender has the option of trying again in a little while. The *A* and *C* bits are present twice in the *Frame status* to increase reliability inasmuch as they are not covered by the checksum.

The *Ending delimiter* contains an *E* bit which is set if any interface detects an error (e.g., a non-Manchester pattern where that is not permitted). It also contains a bit that can be used to mark the last frame in a logical sequence, sort of like an end-of-file bit.

The 802.5 protocol has an elaborate scheme for handling multiple priority frames. The 3-byte token frame contains a field in the middle byte giving the priority of the token. When a station wants to transmit a priority n frame, it must wait until it can capture a token whose priority is less than or equal to n . Furthermore, when a data frame goes by, a station can try to reserve the next token by writing the priority of the frame it wants to send into the frame's *Reservation bits*. However, if a higher priority has already been reserved there, the station may not make a reservation. When the current frame is finished, the next token is generated at the priority that has been reserved.

A little thought will show that this mechanism acts like a ratchet, always jacking the reservation priority higher and higher. To eliminate this problem, the protocol contains some complex rules. The essence of the idea is that the station raising the priority is responsible for lowering the priority again when it is done.

Notice that this priority scheme is substantially different from the token bus scheme, in which each station always gets its fair share of the bandwidth, no

matter what other stations are doing. In the token ring, a station with only low priority frames may starve to death waiting for a low priority token to appear. Clearly, the two committees had different taste when trading off good service for high priority traffic versus fairness to all stations.

Ring Maintenance

The token bus protocol goes to considerable lengths to do ring maintenance in a fully decentralized way. The token ring protocol handles maintenance quite differently. Each token ring has a **monitor station** that oversees the ring. If the monitor goes down, a contention protocol ensures that another station is quickly elected as monitor. (Every station has the capability of becoming the monitor.) While the monitor is functioning properly, it alone is responsible for seeing that the ring operates correctly.

When the ring comes up or any station notices that there is no monitor, it can transmit a CLAIM TOKEN control frame. If this frame circumnavigates the ring before any other CLAIM TOKEN frames are sent, the sender becomes the new monitor (each station has monitor capability built in). The token ring control frames are shown in Fig. 4-31.

Control field	Name	Meaning
00000000	Duplicate address test	Test if two stations have the same address
00000010	Beacon	Used to locate breaks in the ring
00000011	Claim token	Attempt to become monitor
00000100	Purge	Reinitialize the ring
00000101	Active monitor present	Issued periodically by the monitor
00000110	Standby monitor present	Announces the presence of potential monitors

Fig. 4-31. Token ring control frames.

Among the monitor's responsibilities are seeing that the token is not lost, taking action when the ring breaks, cleaning the ring up when garbled frames appear, and watching out for orphan frames. An orphan frame occurs when a station transmits a short frame in its entirety onto a long ring and then crashes or is powered down before the frame can be drained. If nothing is done, the frame will circulate forever.

To check for lost tokens, the monitor has a timer that is set to the longest possible tokenless interval: each station transmitting for the full token-holding time. If this timer goes off, the monitor drains the ring and issues a new token.

When a garbled frame appears, the monitor can detect it by its invalid format or checksum, open the ring to drain it, and issue a new token when the ring has

been cleaned up. Finally, the monitor detects orphan frames by setting the *monitor* bit in the *Access control* byte whenever it passes through. If an incoming frame has this bit set, something is wrong since the same frame has passed the monitor twice without having been drained, so the monitor drains it.

One last monitor function concerns the length of the ring. The token is 24 bits long, which means that the ring must be big enough to hold 24 bits. If the 1-bit delays in the stations plus the cable length add up to less than 24 bits, the monitor inserts extra delay bits so that a token can circulate.

One maintenance function that cannot be handled by the monitor is locating breaks in the ring. When a station notices that either of its neighbors appears to be dead, it transmits a BEACON frame giving the address of the presumably dead station. When the beacon has propagated around as far as it can, it is then possible to see how many stations are down and delete them from the ring using the bypass relays in the wire center, all without human intervention.

It is instructive to compare the approaches taken to controlling the token bus and the token ring. The 802.4 committee was scared to death of having any centralized component that could fail in some unexpected way and take the system down with it. Therefore they designed a system in which the current token holder had special powers (e.g., soliciting bids to join the ring), but no station was otherwise different from the others (e.g., currently assigned administrative responsibility for maintenance).

The 802.5 committee, on the other hand, felt that having a centralized monitor made handling lost tokens, orphan frames and so on much easier. Furthermore, in a normal system, stations hardly ever crash, so occasionally having to put up with contention for a new monitor is not a great hardship. The price paid is that if the monitor ever really goes berserk but continues to issue ACTIVE MONITOR PRESENT control frames periodically, no station will ever challenge it. Monitors cannot be impeached.

This difference in approach comes from the different application areas the two committees had in mind. The 802.4 committee was thinking in terms of factories with large masses of metal moving around under computer control. Network failures could result in severe damage and had to be prevented at all costs. The 802.5 committee was interested in office automation, where a failure once in a rare while could be tolerated as the price for a simpler system. Whether 802.4 is, in fact, more reliable than 802.5 is a matter of some controversy.

4.3.4. Comparison of 802.3, 802.4, and 802.5

With three different and incompatible LANs available, each with different properties, many organizations are faced with the question: Which one should we install? In this section we will look at all three of the 802 LAN standards, pointing out their strengths and weaknesses, comparing and contrasting them.

their detractors.
The major issues
For example, for

net
contains a
er
res subnet
when VC
ickets
e
ssed
ed
inated
uffers
ed in
ch VC

l circuits and
bandwidth. Vir-
tual destination
address in every
hence wasted
the table space
ication circuits

ing virtual cir-
cles. However,
net is easy: the
out where the
is required to
on within the

subnet because resources can be reserved in advance, when the connection is established. Once the packets start arriving, the necessary bandwidth and router capacity will be there. With a datagram subnet, congestion avoidance is more difficult.

For transaction processing systems (e.g., stores calling up to verify credit card purchases), the overhead required to set up and clear a virtual circuit may easily dwarf the use of the circuit. If the majority of the traffic is expected to be of this kind, the use of switched virtual circuits inside the subnet makes little sense. On the other hand, permanent virtual circuits, which are set up manually and last for months or years, may be useful here.

Virtual circuits also have a vulnerability problem. If a router crashes and loses its memory, even if it comes back up a second later, all the virtual circuits passing through it will have to be aborted. In contrast, if a datagram router goes down, only those users whose packets were queued up in the router at the time will suffer, and maybe not even all those, depending upon whether they have already been acknowledged or not. The loss of a communication line is fatal to virtual circuits using it but can be easily compensated for if datagrams are used. Datagrams also allow the routers to balance the traffic throughout the subnet, since routes can be changed halfway through a connection.

It is worth explicitly pointing out that the service offered (connection-oriented or connectionless) is a separate issue from the subnet structure (virtual circuit or datagram). In theory, all four combinations are possible. Obviously, a virtual circuit implementation of a connection-oriented service and a datagram implementation of a connectionless service are reasonable. Implementing connections using datagrams also makes sense when the subnet is trying to provide a highly robust service.

The fourth possibility, a connectionless service on top of a virtual circuit subnet, seems strange but certainly occurs. The obvious example is running IP over an ATM subnet. Here it is desired to run an existing connectionless protocol over a new connection-oriented network layer. As mentioned earlier, this is more of an ad hoc solution to a problem than a good design. In a new system designed to run over an ATM subnet, one would not normally put a connectionless protocol like IP over a connection-oriented network layer like ATM and then layer a connection-oriented transport protocol on top of the connectionless protocol. Examples of all four cases are shown in Fig. 5-3.

5.2. ROUTING ALGORITHMS

The main function of the network layer is routing packets from the source machine to the destination machine. In most subnets, packets will require multiple hops to make the journey. The only notable exception is for broadcast

Upper layer	Type of subnet	
	Datagram	Virtual circuit
Connectionless	UDP over IP	UDP over IP over ATM
Connection-oriented	TCP over IP	ATM AAL1 over ATM

Fig. 5-3. Examples of different combinations of service and subnet structure.

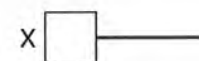
networks, but even here routing is an issue if the source and destination are not on the same network. The algorithms that choose the routes and the data structures that they use are a major area of network layer design.

The **routing algorithm** is that part of the network layer software responsible for deciding which output line an incoming packet should be transmitted on. If the subnet uses datagrams internally, this decision must be made anew for every arriving data packet since the best route may have changed since last time. If the subnet uses virtual circuits internally, routing decisions are made only when a new virtual circuit is being set up. Thereafter, data packets just follow the previously established route. The latter case is sometimes called **session routing**, because a route remains in force for an entire user session (e.g., a login session at a terminal or a file transfer).

Regardless of whether routes are chosen independently for each packet or only when new connections are established, there are certain properties that are desirable in a routing algorithm: correctness, simplicity, robustness, stability, fairness, and optimality. Correctness and simplicity hardly require comment, but the need for robustness may be less obvious at first. Once a major network comes on the air, it may be expected to run continuously for years without systemwide failures. During that period there will be hardware and software failures of all kinds. Hosts, routers, and lines will go up and down repeatedly, and the topology will change many times. The routing algorithm should be able to cope with changes in the topology and traffic without requiring all jobs in all hosts to be aborted and the network to be rebooted every time some router crashes.

Stability is also an important goal for the routing algorithm. There exist routing algorithms that never converge to equilibrium, no matter how long they run. Fairness and optimality may sound obvious—surely no one would oppose them—but as it turns out, they are often contradictory goals. As a simple example of this conflict, look at Fig. 5-4. Suppose that there is enough traffic between A and A' , between B and B' , and between C and C' to saturate the horizontal links. To maximize the total flow, the X to X' traffic should be shut off altogether.

Unfortunately,
between global



Before we
ity, we must d
delay is an ob
Furthermore, tl
system near cap
works attempt
reducing the n
amount of band

Routing alg
adaptive. Non
urements or es
the route to use
and download
sometimes call

Adaptive :
changes in the
differ in where
from all router
changes, or wh
(e.g., distance,
tions we will di

5.2.1. The Op

Before gett
make a genera
ogy or traffic.
if router J is on

Unfortunately, X and X' may not see it that way. Evidently, some compromise between global efficiency and fairness to individual connections is needed.

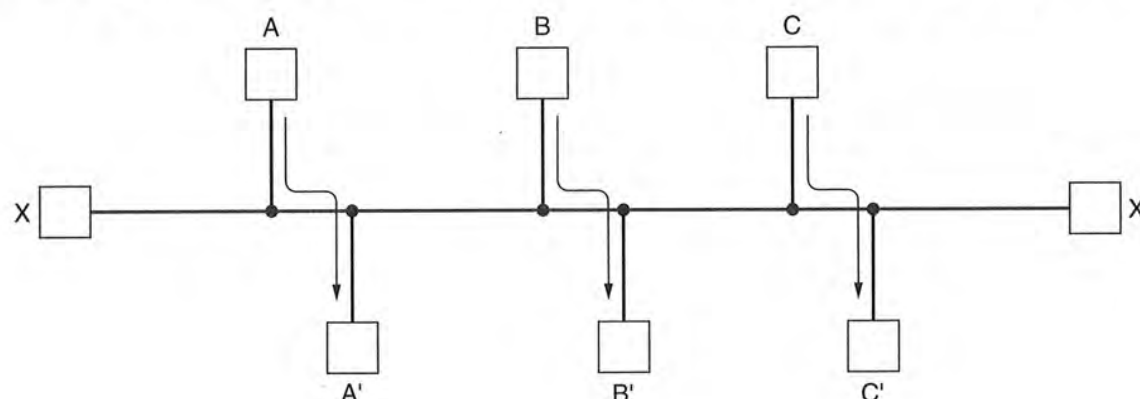


Fig. 5-4. Conflict between fairness and optimality.

Before we can even attempt to find trade-offs between fairness and optimality, we must decide what it is we seek to optimize. Minimizing mean packet delay is an obvious candidate, but so is maximizing total network throughput. Furthermore, these two goals are also in conflict, since operating any queueing system near capacity implies a long queueing delay. As a compromise, many networks attempt to minimize the number of hops a packet must make, because reducing the number of hops tends to improve the delay and also reduce the amount of bandwidth consumed, which tends to improve the throughput as well.

Routing algorithms can be grouped into two major classes: nonadaptive and adaptive. **Nonadaptive algorithms** do not base their routing decisions on measurements or estimates of the current traffic and topology. Instead, the choice of the route to use to get from I to J (for all I and J) is computed in advance, off-line, and downloaded to the routers when the network is booted. This procedure is sometimes called **static routing**.

Adaptive algorithms, in contrast, change their routing decisions to reflect changes in the topology, and usually the traffic as well. Adaptive algorithms differ in where they get their information (e.g., locally, from adjacent routers, or from all routers), when they change the routes (e.g., every ΔT sec, when the load changes, or when the topology changes), and what metric is used for optimization (e.g., distance, number of hops, or estimated transit time). In the following sections we will discuss a variety of routing algorithms, both static and dynamic.

5.2.1. The Optimality Principle

Before getting into specific algorithms, it may be helpful to note that one can make a general statement about optimal routes without regard to network topology or traffic. This statement is known as the **optimality principle**. It states that if router J is on the optimal path from router I to router K , then the optimal path

from J to K also falls along the same route. To see this, call the part of the route from I to J r_1 and the rest of the route r_2 . If a route better than r_2 existed from J to K , it could be concatenated with r_1 to improve the route from I to K , contradicting our statement that $r_1 r_2$ is optimal.

As a direct consequence of the optimality principle, we can see that the set of optimal routes from all sources to a given destination form a tree rooted at the destination. Such a tree is called a **sink tree** and is illustrated in Fig. 5-5 where the distance metric is the number of hops. Note that a sink tree is not necessarily unique; other trees with the same path lengths may exist. The goal of all routing algorithms is to discover and use the sink trees for all routers.

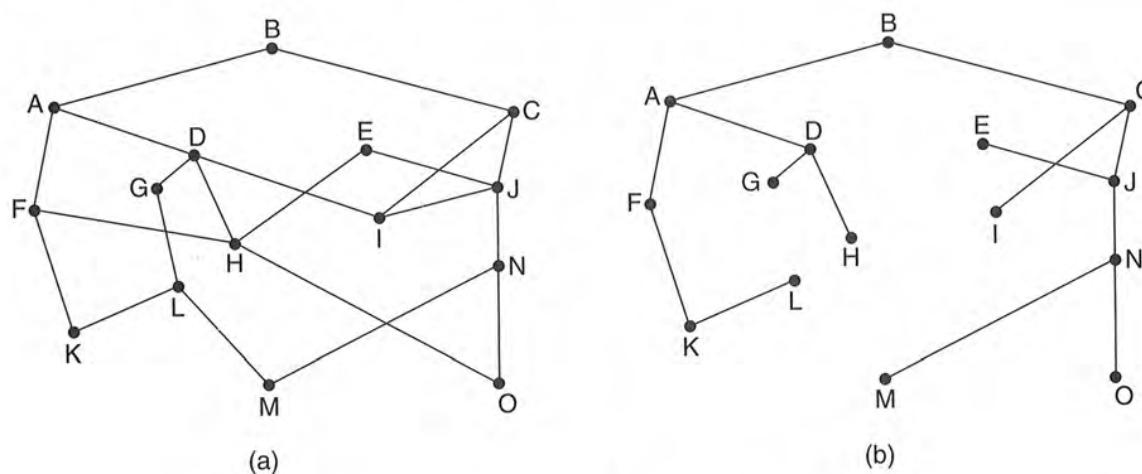


Fig. 5-5. (a) A subnet. (b) A sink tree for router B.

Since a sink tree is indeed a tree, it does not contain any loops, so each packet will be delivered within a finite and bounded number of hops. In practice, life is not quite this easy. Links and routers can go down and come back up during operation, so different routers may have different ideas about the current topology. Also, we have quietly finessed the issue of whether each router has to individually acquire the information on which to base its sink tree computation, or whether this information is collected by some other means. We will come back to these issues shortly. Nevertheless, the optimality principle and the sink tree provide a benchmark against which other routing algorithms can be measured.

In the next three sections, we will look at three different static routing algorithms. After that we will move on to adaptive ones.

5.2.2. Shortest Path Routing

Let us begin our study of routing algorithms with a technique that is widely used in many forms because it is simple and easy to understand. The idea is to build a graph of the subnet, with each node of the graph representing a router and

each arc of the choose a route est path between

The conce measuring pat and ABE in Fi in kilometers, figure is draw

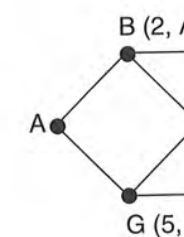
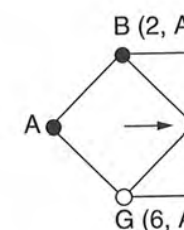
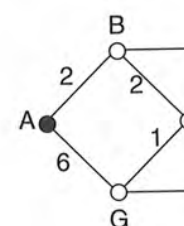


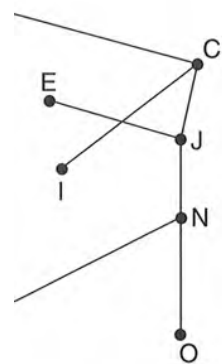
Fig. 5
The a

However tance. For a transmission runs. With the path with the

In the m tion of the d length, meas

rt of the route
existed from J
to K , contra-

that the set of
rooted at the
Fig. 5-5 where
not necessarily
of all routing



so each packet
practice, life is
ack up during
current topol-
has to indivi-
omputation, or
come back to
sink tree pro-
asured.
routing algo-

that is widely
The idea is to
g a router and

each arc of the graph representing a communication line (often called a link). To choose a route between a given pair of routers, the algorithm just finds the shortest path between them on the graph.

The concept of a **shortest path** deserves some explanation. One way of measuring path length is the number of hops. Using this metric, the paths ABC and ABE in Fig. 5-6 are equally long. Another metric is the geographic distance in kilometers, in which case ABC is clearly much longer than ABE (assuming the figure is drawn to scale).

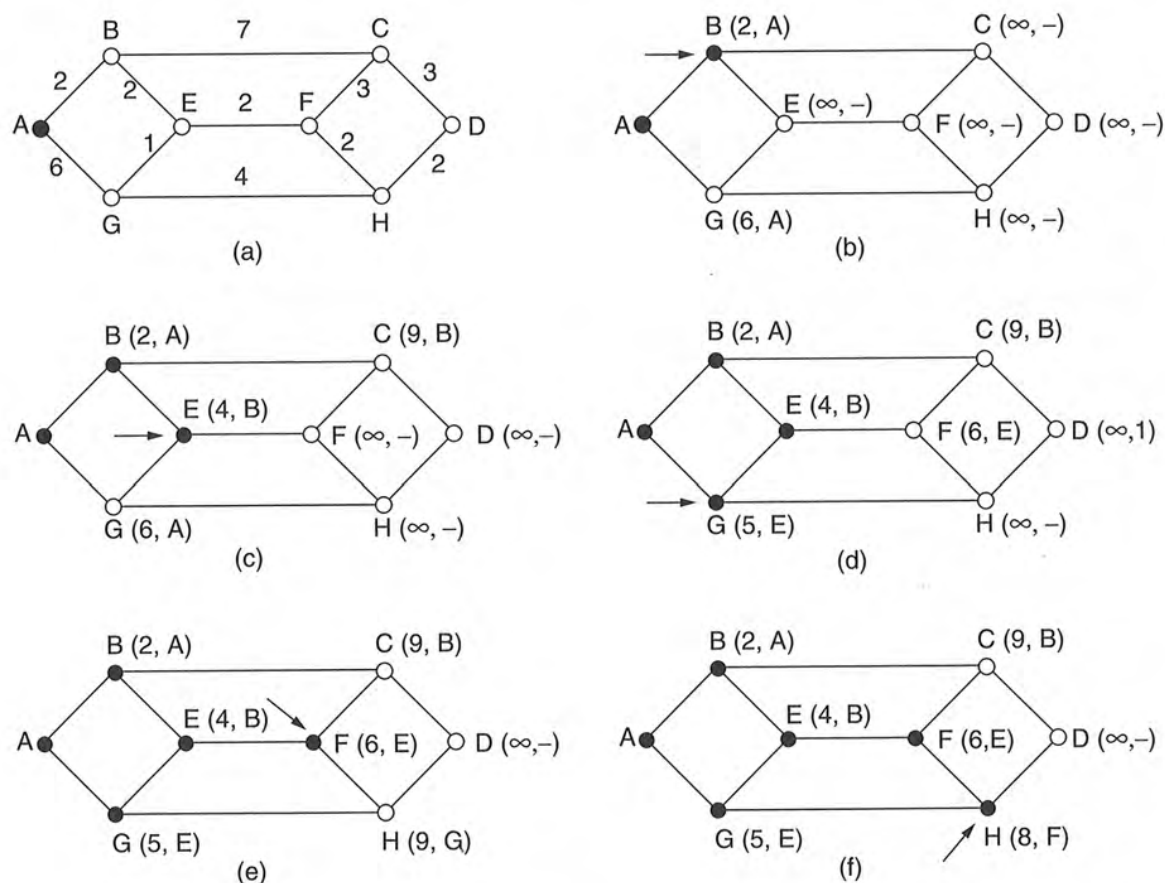


Fig. 5-6. The first five steps used in computing the shortest path from A to D . The arrows indicate the working node.

However, many other metrics are also possible besides hops and physical distance. For example, each arc could be labeled with the mean queueing and transmission delay for some standard test packet as determined by hourly test runs. With this graph labeling, the shortest path is the fastest path, rather than the path with the fewest arcs or kilometers.

In the most general case, the labels on the arcs could be computed as a function of the distance, bandwidth, average traffic, communication cost, mean queue length, measured delay, and other factors. By changing the weighting function,

the algorithm would then compute the "shortest" path measured according to any one of a number of criteria, or a combination of criteria.

Several algorithms for computing the shortest path between two nodes of a graph are known. This one is due to Dijkstra (1959). Each node is labeled (in parentheses) with its distance from the source node along the best known path. Initially, no paths are known, so all nodes are labeled with infinity. As the algorithm proceeds and paths are found, the labels may change, reflecting better paths. A label may be either tentative or permanent. Initially, all labels are tentative. When it is discovered that a label represents the shortest possible path from the source to that node, it is made permanent and never changed thereafter.

To illustrate how the labeling algorithm works, look at the weighted, undirected graph of Fig. 5-6(a), where the weights represent, for example, distance. We want to find the shortest path from *A* to *D*. We start out by marking node *A* as permanent, indicated by a filled in circle. Then we examine, in turn, each of the nodes adjacent to *A* (the working node), relabeling each one with the distance to *A*. Whenever a node is relabeled, we also label it with the node from which the probe was made, so we can reconstruct the final path later. Having examined each of the nodes adjacent to *A*, we examine all the tentatively labeled nodes in the whole graph and make the one with the smallest label permanent, as shown in Fig. 5-6(b). This one becomes the new working node.

We now start at *B*, and examine all nodes adjacent to it. If the sum of the label on *B* and the distance from *B* to the node being considered is less than the label on that node, we have a shorter path, so the node is relabeled.

After all the nodes adjacent to the working node have been inspected and the tentative labels changed if possible, the entire graph is searched for the tentatively labeled node with the smallest value. This node is made permanent and becomes the working node for the next round. Figure 5-6 shows the first five steps of the algorithm.

To see why the algorithm works, look at Fig. 5-6(c). At that point we have just made *E* permanent. Suppose that there were a shorter path than *ABE*, say *AXYZE*. There are two possibilities: either node *Z* has already been made permanent, or it has not been. If it has, then *E* has already been probed (on the round following the one when *Z* was made permanent), so the *AXYZE* path has not escaped our attention.

Now consider the case where *Z* is still tentatively labeled. Either the label at *Z* is greater than or equal to that at *E*, in which case *AXYZE* cannot be a shorter path than *ABE*, or it is less than that of *E*, in which case *Z* and not *E* will become permanent first, allowing *E* to be probed from *Z*.

This algorithm is given in Fig. 5-7. The only difference between the program and the algorithm described above is that in Fig. 5-7, we compute the shortest path starting at the terminal node, *t*, rather than at the source node, *s*. Since the shortest path from *t* to *s* in an undirected graph is the same as the shortest path from *s* to *t*, it does not matter at which end we begin (unless there are several shortest paths,

in which case searching back to the predecessor is thus reversed. The path produced in the

5.2.3. Flooding

Another search algorithm is flooding, in which every node in the network generates vast numbers of copies of the packet. The measures are taken to ensure that the packet counter contains the smallest value with the packet. The packet counter should be incremented. If the sender receives a packet with a smaller counter value, the worst case is that the packet is discarded.

An alternative to flooding is to have packets have a label. To achieve this, the sender of a packet it receives tells which node it came from. If an incoming packet has a smaller label, it is discarded.

To prevent flooding, the network is partitioned by a set of nodes. When a packet is received, it is discarded if it is not in the set. This is a summary of the algorithm.

A variation of this algorithm is to use only those nodes that are usually little used. This topology is called a sparse topology.

Flooding is a search algorithm. For example, in a network with *n* nodes, it takes *n* bits at any one time to distributed databases concurrently. Flooding is as fast as flooding always in parallel. C is ignored the over

ording to any

o nodes of a
s labeled (in
known path.

As the algo-
better paths.
are tentative.
path from the
r.

ie weighted,
example, dis-
t by marking
nine, in turn,
one with the
re node from
ater. Having
ively labeled
permanent, as

e sum of the
less than the

ected and the
he tentatively
and becomes
e steps of the

oint we have
ian *ABE*, say
en made per-
(on the round
path has not

er the label at
t be a shorter
? will become

n the program
: shortest path
e the shortest
th from *s* to *t*,
hortest paths,

in which case reversing the search might discover a different one). The reason for searching backward is that each node is labeled with its predecessor rather than its successor. When copying the final path into the output variable, *path*, the path is thus reversed. By reversing the search, the two effects cancel, and the answer is produced in the correct order.

5.2.3. Flooding

Another static algorithm is **flooding**, in which every incoming packet is sent out on every outgoing line except the one it arrived on. Flooding obviously generates vast numbers of duplicate packets, in fact, an infinite number unless some measures are taken to damp the process. One such measure is to have a hop counter contained in the header of each packet, which is decremented at each hop, with the packet being discarded when the counter reaches zero. Ideally, the hop counter should be initialized to the length of the path from source to destination. If the sender does not know how long the path is, it can initialize the counter to the worst case, namely, the full diameter of the subnet.

An alternative technique for damming the flood is to keep track of which packets have been flooded, to avoid sending them out a second time. One way to achieve this goal is to have the source router put a sequence number in each packet it receives from its hosts. Each router then needs a list per source router telling which sequence numbers originating at that source have already been seen. If an incoming packet is on the list, it is not flooded.

To prevent the list from growing without bound, each list should be augmented by a counter, *k*, meaning that all sequence numbers through *k* have been seen. When a packet comes in, it is easy to check if the packet is a duplicate; if so, it is discarded. Furthermore, the full list below *k* is not needed, since *k* effectively summarizes it.

A variation of flooding that is slightly more practical is **selective flooding**. In this algorithm the routers do not send every incoming packet out on every line, only on those lines that are going approximately in the right direction. There is usually little point in sending a westbound packet on an eastbound line unless the topology is extremely peculiar.

Flooding is not practical in most applications, but it does have some uses. For example, in military applications, where large numbers of routers may be blown to bits at any instant, the tremendous robustness of flooding is highly desirable. In distributed database applications, it is sometimes necessary to update all the databases concurrently, in which case flooding can be useful. A third possible use of flooding is as a metric against which other routing algorithms can be compared. Flooding always chooses the shortest path, because it chooses every possible path in parallel. Consequently, no other algorithm can produce a shorter delay (if we ignore the overhead generated by the flooding process itself).


```

#define MAX_NODES 1024          /* maximum number of nodes */
#define INFINITY 1000000000     /* a number larger than every maximum path */
int n, dist[MAX_NODES][MAX_NODES]; /* dist[i][j] is the distance from i to j */

void shortest_path(int s, int t, int path[])
{ struct state {                /* the path being worked on */
    int predecessor;            /* previous node */
    int length;                 /* length from source to this node */
    enum {permanent, tentative} label; /* label state */
} state[MAX_NODES];

int i, k, min;
struct state *p;
for (p = &state[0]; p < &state[n]; p++) { /* initialize state */
    p->predecessor = -1;
    p->length = INFINITY;
    p->label = tentative;
}
state[t].length = 0; state[t].label = permanent;
k = t; /* k is the initial working node */
do { /* Is there a better path from k? */
    for (i = 0; i < n; i++) /* this graph has n nodes */
        if (dist[k][i] != 0 && state[i].label == tentative) {
            if (state[k].length + dist[k][i] < state[i].length) {
                state[i].predecessor = k;
                state[i].length = state[k].length + dist[k][i];
            }
        }

    /* Find the tentatively labeled node with the smallest label. */
    k = 0; min = INFINITY;
    for (i = 0; i < n; i++)
        if (state[i].label == tentative && state[i].length < min) {
            min = state[i].length;
            k = i;
        }
    state[k].label = permanent;
} while (k != s);

/* Copy the path into the output array. */
i = 0; k = s;
do {path[i++] = k; k = state[k].predecessor; } while (k >= 0);
}

```

Fig. 5-7. Dijkstra's algorithm to compute the shortest path through a graph.

5.2.4. Flow-I

The algorithm does not consider the path from A to B, AGEFC, even though it is a static flow-based routing.

In some routing algorithms, the path is stable and predictable. In each step, a well-defined total volume of traffic is sent, and the average traffic is constant. The routine is to initialize the routing.

The basic algorithm for average flow is to find a line from queue to queue, and then to calculate the whole subnet that produces the flow.

To use the algorithm, the subnet to be used is given. Third, the path must be chosen.

As an example, the weights are 3 kbps. The maximum entry for source is also the number of packets, for example, 3. Notice that the path is shown in the figure.

Given the path, the flow is shown. For example, the packets/sec for each of the three columns of flow is identical. The traffic is identical for each line, and the

5.2.4. Flow-Based Routing

The algorithms studied so far take only the topology into account. They do not consider the load. If, for example, there is always a huge amount of traffic from A to B , in Fig. 5-6, then it may be better to route traffic from A to C via $AGEFC$, even though this path is much longer than ABC . In this section we will study a static algorithm that uses both topology and load for routing. It is called **flow-based routing**.

In some networks, the mean data flow between each pair of nodes is relatively stable and predictable. For example, in a corporate network for a retail store chain, each store might send orders, sales reports, inventory updates, and other well-defined types of messages to known sites in a predefined pattern, so that the total volume of traffic varies little from day to day. Under conditions in which the average traffic from i to j is known in advance and, to a reasonable approximation, constant in time, it is possible to analyze the flows mathematically to optimize the routing.

The basic idea behind the analysis is that for a given line, if the capacity and average flow are known, it is possible to compute the mean packet delay on that line from queueing theory. From the mean delays on all the lines, it is straightforward to calculate a flow-weighted average to get the mean packet delay for the whole subnet. The routing problem then reduces to finding the routing algorithm that produces the minimum average delay for the subnet.

To use this technique, certain information must be known in advance. First the subnet topology must be known. Second, the traffic matrix, F_{ij} , must be given. Third, the line capacity matrix, C_{ij} , specifying the capacity of each line in bps must be available. Finally, a (possibly tentative) routing algorithm must be chosen.

As an example of this method, consider the full-duplex subnet of Fig. 5-8(a). The weights on the arcs give the capacities, C_{ij} , in each direction measured in kbps. The matrix of Fig. 5-8(b) has an entry for each source-destination pair. The entry for source i to destination j shows the route to be used for i - j traffic, and also the number of packets/sec to be sent from source i to destination j . For example, 3 packets/sec go from B to D , and they use route BFD to get there. Notice that some routing algorithm has already been applied to derive the routes shown in the matrix.

Given this information, it is straightforward to calculate the total in line i , λ_i . For example, the B - D traffic contributes 3 packets/sec to the BF line and also 3 packets/sec to the FD line. Similarly, the A - D traffic contributes 1 packet/sec to each of three lines. The total traffic in each eastbound line is shown in the λ_i column of Fig. 5-9. In this example, all the traffic is symmetric, that is, the XY traffic is identical to the YX traffic, for all X and Y . In real networks this condition does not always hold. The figure also shows the mean number of packets/sec on each line, μC_i assuming a mean packet size of $1/\mu = 800$ bits.

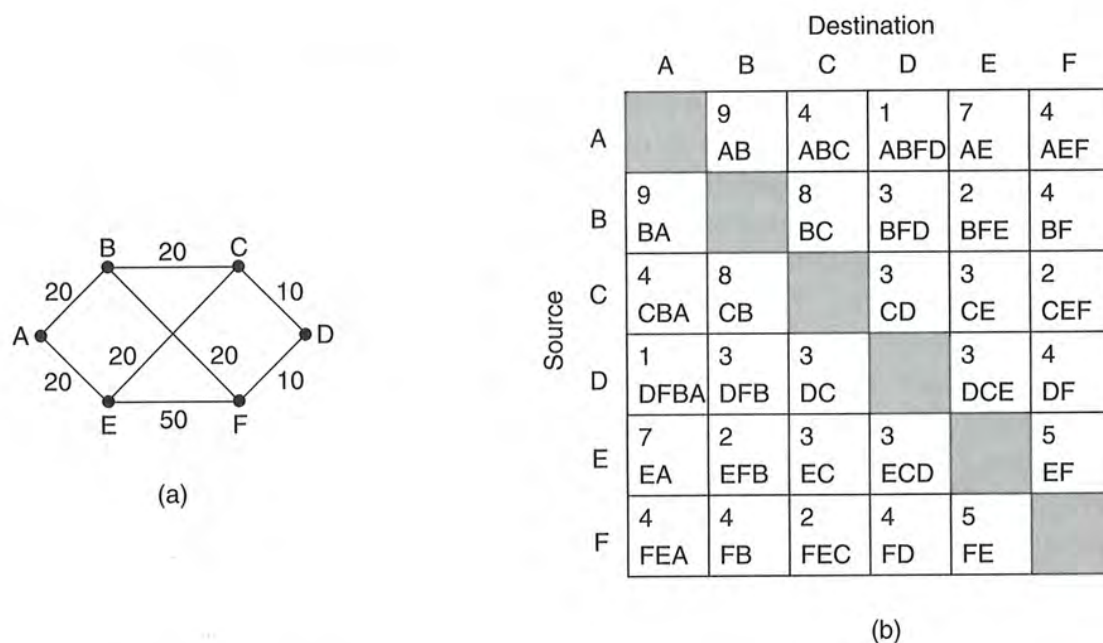


Fig. 5-8. (a) A subnet with line capacities shown in kbps. (b) The traffic in packets/sec and the routing matrix.

The next-to-last column of Fig. 5-9 gives the mean delay for each line derived from the queueing theory formula

$$T = \frac{1}{\mu C - \lambda}$$

where $1/\mu$ is the mean packet size in bits, C is the capacity in bps, and λ is the mean flow in packets/sec. For example, with a capacity $\mu C = 25$ packets/sec and an actual flow $\lambda = 14$ packets/sec, the mean delay is 91 msec. Note that with $\lambda = 0$, the mean delay is still 40 msec, because the capacity is 25 packets/sec. In other words, the "delay" includes both queueing and service time.

To compute the mean delay time for the entire subnet, we take the weighted sum of each of the eight lines, with the weight being the fraction of the total traffic using that line. In this example, the mean turns out to be 86 msec.

To evaluate a different routing algorithm, we can repeat the entire process, only with different flows to get a new average delay. If we restrict ourselves to only single path routing algorithms, as we have done so far, there are only a finite number of ways to route packets from each source to each destination. It is always possible to write a program to simply try them all, one after another, and find out which one has the smallest mean delay. Since this calculation can be done off-line in advance, the fact that it may be time consuming is not necessarily a serious problem. This one is then the best routing algorithm. Bertsekas and Gallager (1992) discuss flow-based routing in detail.

i	Line
1	AB
2	BC
3	CD
4	AE
5	EF
6	FD
7	BF
8	EC

Fig. 5-
The rev

5.2.5. Distance

Modern c
than the static
tance vector r
we will look a
one.

Distance
table (i.e., a v
line to use to
the neighbors.

The dista
including the
Fulkerson alg
Ford and Full
was also used
and Novell's
protocols.

In distanc
and containin
parts: the pref
time or distar
time delay in
something sin

The route
metric is hop
router simply