Introduction to Computer Networks

(Source: A.S.Tanenbaum: Computer Networks, 3rd. edition)

Doc. RNDr. Peter Mederly, CSc.

Editors:

Andrej Jursa 4jursa@st.fmph.uniba.sk
Jozef Fulop 4fulop@st.fmph.uniba.sk
Lubor Illek 4illek@st.fmph.uniba.sk

Contents

1. Introduction

Each of the past three centuries has been dominated by a single technology:

18th - industrial revolution with the great mechanical systems

19th - age of steam engine

20th - the key technology has been information gathering, processing and distribution - telephones, radio, television, computer industry, communication satellites.

These areas are rapidly converging.

The ability to process information grows - the demand for even more sophisticated information processing grows even faster.

Progress of computers:

- First two decades - highly centralized systems.
- Later - the merging of computers and communications has had profound influence on the way computer systems are organized - replacement of the old model of highly centralized systems by computer networks.

Computer network is an interconnected collection of autonomous computers.

Two computers are said to be *interconnected* if they are able to exchange information. The connection can be realized by different media.

Autonomous means no master/slave like. A system with one control unit and many slaves is not a network; nor is a large computer with remote printers and terminals.

Computer networks vs. *distributed systems*:

- In distributed systems the existence of multiple autonomous computers is transparent to the user - the system looks like a virtual uniprocessor.

With a network, users must explicitly log onto one machine.

- Distributed system is a software system built on top of a network.

Another definition of a distributed system: interconnected collection of autonomous computers, processes. The computers, processes, or processors are referred to as the nodes of the distributed system.

1.1. Uses of computer networks

Goals of the networks for companies:

- Resource sharing - programs, data, equipment.
- High reliability - replicated files, multiple CPU.

- Saving money - small computers have much better price/performance ratio than large ones. The systems of personal computers, one per person, are built with data kept on one or more shared file server machines. Users are called clients, the whole arrangement is called the client-server model.
- Scalability - the ability to increase system performance gradually as the workload grows just by adding more processors.
- Communication medium - enables e.g. to write a report together.

In long run, the use of networks to enhance human-to-human communication will probably prove more important than technical goals such as improved reliability.

Services delivered by networks to private individuals at home:

- Access to remote information (interaction between a person and a remote database) - financial institutions, home shopping, newspapers, digital library, potential replacement of printed books by notebook computers, access to information systems (WWW).
- Person-to-person communication (21st century answer to the 19th century's telephone) - email, videoconference, newsgroups.
- Interactive entertainment - video on demand, interactive films.

The widespread introduction of networking will introduce new social, ethical, political problems forming social issues of networking, e.g.:

- newsgroups set up on topics that people actually care about (politics, religion, sex) - photographs, videoclips (e.g.children pornography)
- employee rights versus employer rights - some employers have claimed the right to read and possibly censor employee messages
- school and students
- anonymous messages

Computer networks, like the printing press 500 years ago, allow ordinary citizens to distribute their views in different ways and to different audiences than were previously possible. This new-found freedom brings with it many unsolved social, political, and moral issues.

1.2. Network hardware

There is no generally accepted taxonomy into which all computer networks fit, but two dimensions stand out as important: transmission technology and scale.

Classification of networks according to transmission technology:

- broadcast networks,
- point-to-point networks.

Broadcast networks are networks with single communication channel shared by all the machines. Short messages (packets) sent by any machine are received by all others. An address field within the packet specifies for whom it is intended. Analogy: someone shout in the corridor with many rooms.

Broadcasting is a mode of operation in which a packet is sent to every machine using a special code in the address field.

Multicasting is sending a packet to a subset of the machines.

Point-to-point networks consist of many connections between individual pairs of machines. In these types of networks:

- A packet on its way from the source to the destination may go through intermediate machines.
- In general, multiple routes are possible - routing algorithms are necessary.

General rule (with many exceptions): smaller, geographically localized networks tends to use broadcasting, larger networks usually are point-to-point.

Classification of networks by scale: If we take as a criterion the interprocessor distance, we get on the one side of the scale data flow machines, highly parallel computers with many functional units all working on the same program. Next come the multicomputers, systems that communicate through short, very fast buses. Beyond the multicomputers are the true networks, computers communicating over longer cables. Finally, the connection of two or more networks is called an internetwork. Distance is important as a classification metric because different techniques are used at different scales.

1.2.1. Local Area Networks

Local area networks (LANs) re privately-owned, within a single building or campus, of up to a few kilometers in size. They are distinguished from other kind of networks by three characteristics:

- size,
- transmission technology,
- topology.

LANs are restricted in size - the worst-case transmission time is known in advance, it makes possible to use certain kinds of design.

LANs transmission technology often consists of a single cable to which all machines are attached. Traditional LANs run at speed of 10 to 100 Mbps. Newer LANs may operate at higher speeds.

Possible topologies for broadcast LANs (Fig. 1-3):

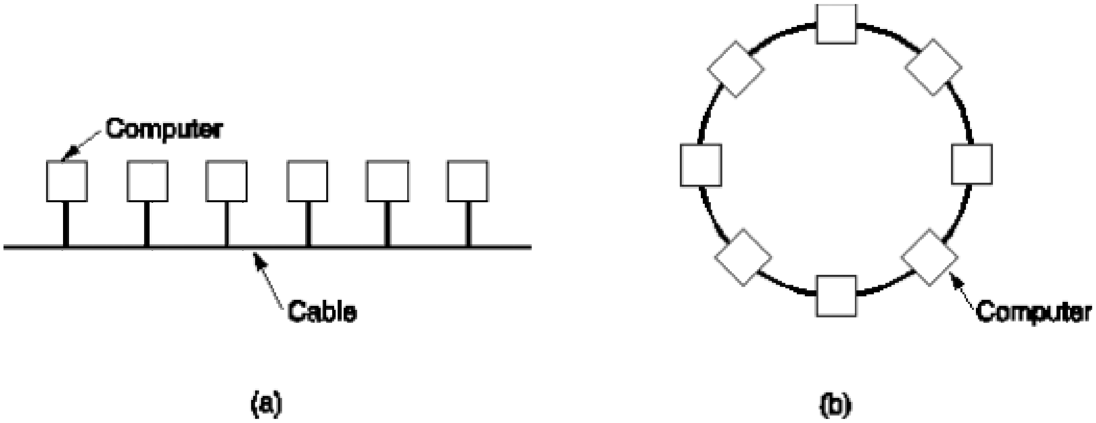


Fig. 1-3. Two broadcast networks. (a) Bus. (b) Ring.

- *bus* - at any instant one machine is the master of the bus allowed to transmit. Arbitration mechanism for resolving the conflicts when more than one machine want to transmit may be centralized or distributed. Example: Ethernet as a bus-based broadcast network with decentralized control operating at 10 or 100 Mbps.

- *ring* - each bit propagates around, typically it circumnavigates the entire ring in the time it takes to transmit a few bits, often before the complete packet has even be transmitted. Example: IBM token ring operating at 4 and 16 Mbps.

Broadcast networks can be, depending on how the channel is allocated, further divided into:

- Static - a typical would be a time division for the access to the channel and round-robin algorithms. It wastes channel capacity.
- Dynamic - on demand. Channel allocation could be centralized or decentralized.

LAN built using point-to-point lines is really a miniature WAN.

1.2.2. Metropolitan Area Networks

Metropolitan area network (MAN) is basically a bigger version of a LAN and normally uses similar technology. It might cover a group of nearby corporate offices or a city and might be either private or public. The main reason for even distinguishing MANs as a special category is that a standard has been adopted for them. It is called DQDB (Distributed Queue Dual Bus).

1.2.3. Wide Area Networks

A *wide area network* (WAN):

- spans a large geographical area,
- contains hosts (or end-systems) intended for running user programs,
- the hosts are connected by a *subnet* that carries messages from host to host.

The subnet usually consists of transmission lines (circuits, channels, or trunks) and switching elements. The switching elements are specialized computers used to connect two or more transmission lines. There is no standard technology used to name switching elements (e.g. packet switching nodes, intermediate systems, data switching exchanges). As a generic term we will use the word router. (Fig. 1-5)

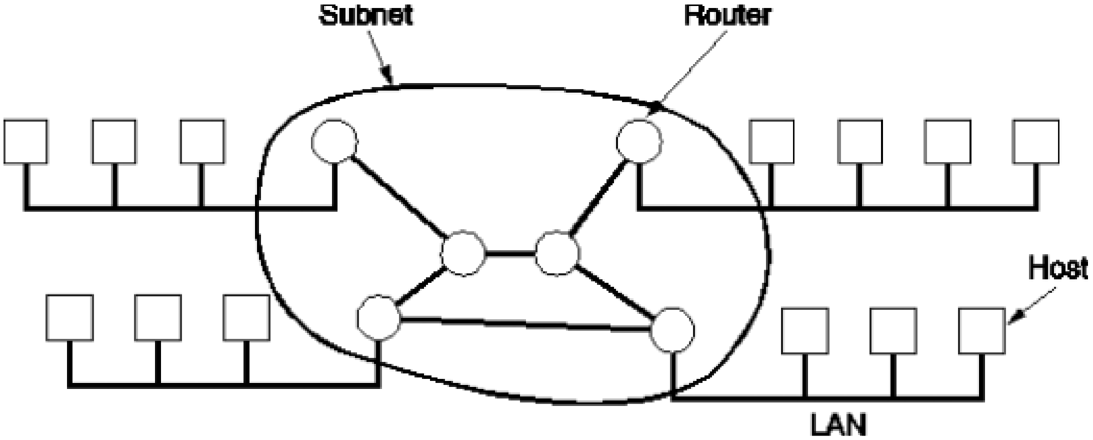


Fig. 1-5. Relation between hosts and the subnet.

Remark: the term "subnet" also acquired a second meaning in conjuction with network addressing.

If two routers that do not share a cable wish to communicate, they must do it via other routers. When a packet is sent from one router to another via intermediate routers, the packet is received at each intermediate router, stored there until the required output line is free, and then forwarded. A subnet using this principle is called point-to-point, store-and-forward, or packet-switched subnet. Nearly all wide area networks (except those using satellites) have store-and-forward subnets.

When the packets are small and all the same size, they are often called cells.

A second possibility for a WAN is a satellite or ground radio system. Each router has an antenna through which it can send and receive. All router can hear the output from the satellite. Satellite networks are inherently broadcast.

1.2.4. Wireless Networks

The owners of mobile computers want to be connected to their home base when they are away from home. In case where wired connection is impossible (in cars, airplanes), the wireless networks are necessary.

The use of wireless networks:

- portable office - sending and receiving telephone calls, faxes, e-mails, remote login, ...
- rescue works,
- keeping in contact,
- military.

Wireless networking and *mobile computing* are often related but they are not identical. Portable computers are sometimes wired (e.g. at the traveler's stay in a hotel) and some wireless computer are not portable (e.g. in the old building without any network infrastructure).

Wireless LANs are easy to install but they have also some disadvantages: lower capacity (1-2 Mbps, higher error rate, possible interference of the transmissions from different computers).

Wireless networks come in many forms:

- antennas all over the campus to allow to communicate from under the trees,
- using a cellular (i.e. portable) telephone with a traditional analog modem,
- direct digital cellular service called CDPD (Cellular Digital Packet Data),
- different combinations of wired and wireless networking.

1.2.5. Internetworks

Internetwork or *internet* is a collection of interconnected networks. A common form of internet is a collection of LAN connected by WAN. Connecting incompatible networks together requires using machines called gateways to provide the necessary translation.

Internet (with uppercase I) means a specific worldwide internet.

Subnets, networks and internetworks are often confused.

Subnet makes the most sense in the context of a wide area network, where it refers to the collection of routers and communication lines. The combination of a subnet and its hosts forms a network. An internetwork is formed when distinct networks are connected together.

1.3. Network software

The first networks were designed with the hardware as the main concern and the software afterthought. This strategy no longer works. Network software is now highly structured.

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 actual structure of layers differs from network to network.

Layer n on one machine carries on a conversation with layer n on another machine. The rules and conventions used in this conversation are known as layer n protocol (Fig. 1-9).

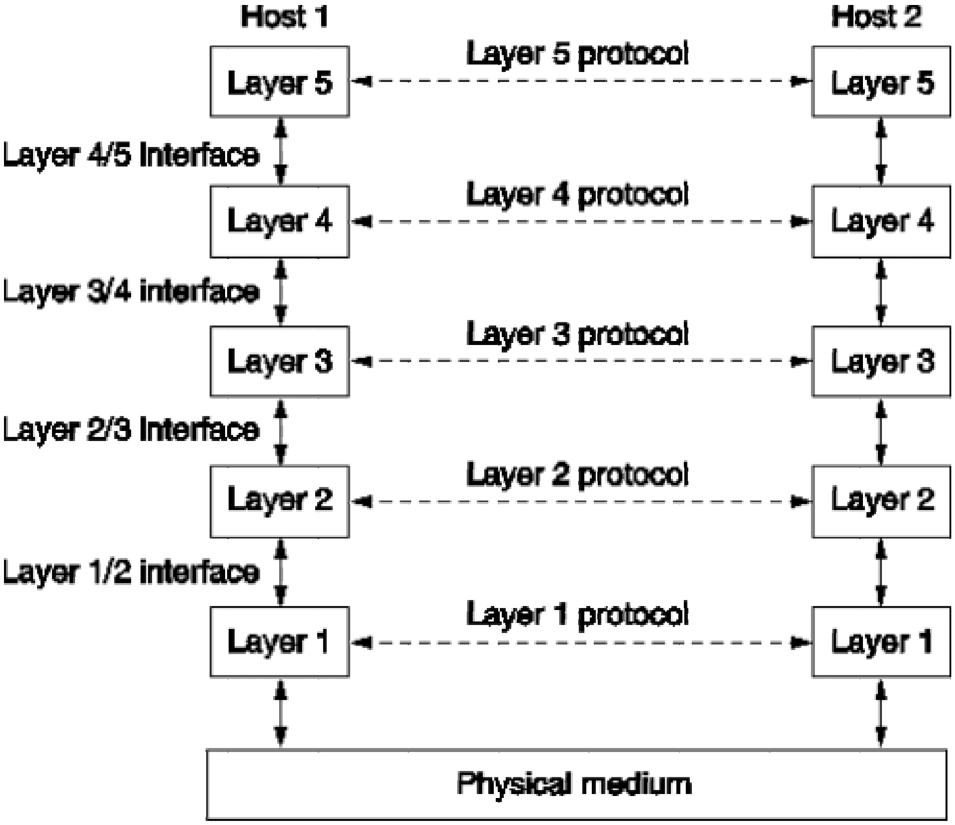


Fig. 1-9. Layers, protocols, and interfaces.

Data between layers n on different machines are not transferred directly. Each layer passes data and control information to the layer directly below it until the lowest layer is reached. Below layer 1 there is a physical medium through which actual communication occurs.

Between each pair of adjacent layers there is an *interface*. The interface defines which primitive operations and services the lower layer offers to the upper one.

A set of layers and protocols is called network architecture. A list of protocols used by a certain system, one protocol per layer, is called a *protocol stack*.

The *peer process* abstraction is crucial to all network design. Using it, the unmanageable task of designing the complete network can be broken into several smaller, manageable, design problems, namely the design of the individual layers.

Lower layers of the protocol hierarchy are frequently implemented in hardware or firmware.

1.3.1. Design Issues for the Layers

Some of the key design issues that cur in computer networking are present in several layers. The more important ones are:

- a mechanism for identifying senders and receivers - addressing,
- the rules for data transfer - simplex, half-duplex, full-duplex communication,
- error control - error-detecting and error-correcting codes,
- sequencing of messages,
- the problem of fast sender and slow receiver,
- inability to accept arbitrarily long messages,
- the effective transmission of small messages,
- multiplexing and demultiplexing,
- routing.

1.3.2. Interfaces and Services

The function of each layer is to provide services to the layer above it. What a service is in more detail?

1.3.3. Some terminology

The active elements in each layer are called *entities*. An entity can be a software entity (such as a process), or a hardware entity (such as an intelligent I/O chip). Entities in the same layer on different machines are called peer entities.

The entities in layer n implement a service used by layer n+1. Layer n is the service provider for the layer n+1 being the service user. Layer n may use the services of layer n - 1 in order to provide its service.

Services are available at SAPs (*Service Access Points*). The layer n SAPs are the places, where layer n+1 can access the services offered. Each SAP has an address that uniquely identifies it.

At a typical interface, the layer n+1 entity passes an IDU (Interface Data Unit) to the layer n entity through the SAP. The IDU consists of an SDU (Service Data Unit) and some control information. The SDU is the information passed across the network to the peer entity and then up to layer n+1. The control information is needed to help the lower layer do its job (e.g. the number of bytes in the SDU) but is not part of the data itself.

In order to transfer SDU, the layer n entity may have to fragment it into several pieces, each of which is given a header and sent as a separate PDU (Protocol Data Unit) such as a packet. The PDU headers are used by the peer entities to carry out their peer protocol. They identify which PDU contain data and which contain control information, provide sequence numbers and counts, and so on.

1.3.4. Connection-oriented and Connectionless Services

Layers can offer two different types of service to the layers above them: connection-oriented and connectionless.

Connection-oriented service (modeled after the telephone system): to use it, the service user first establishes a connection, uses the connection, and then releases the connection. The essential aspect of a connection is that it acts like a tube: the sender pushes objects (bits) in at one end, and the receiver takes them out in the same order at the other end.

Connectionless service (modeled after the postal system): Each message carries the full destination address, and each one is routed through the system independent of all the others.

Quality of service - some services are reliable in the sense that they never lose data. Reliability is usually implemented by having the receiver acknowledge the receipt of each message. The acknowledgment process is often worth but introduces sometimes undesirable overheads and delays.

Reliable connection-oriented service has two minor variation:

- message sequences - the message boundaries are preserved.
- byte streams - the connection is simply a stream of bytes, with no message boundaries.

Applications where delays introduced by acknowledgment are unacceptable:

- digitized voice traffic,
- video film transmission.

The use of connectionless services:

- electronic junk mail (third class mail as advertisements) - this service is moreover unreliable (meaning not acknowledged). Such connectionless services are often called datagram services.
- acknowledged datagram services - connectionless datagram services with acknowledgment.
- request-reply service - the sender transmits a single datagram containing a request. The reply contains the answer. Request-reply is commonly used to implement communication in the client-server model.

1.4. Reference models

1.4.1. The OSI Reference Model

The *OSI model* is based on a proposal develop by ISO as a first step toward international standardization of the protocols used in the various layers. The model is called ISO OSI (Open Systems Interconnection) Reference Model.

Open system is a system open for communication with other systems.

The OSI model has 7 layers (Fig. 1-16). 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.

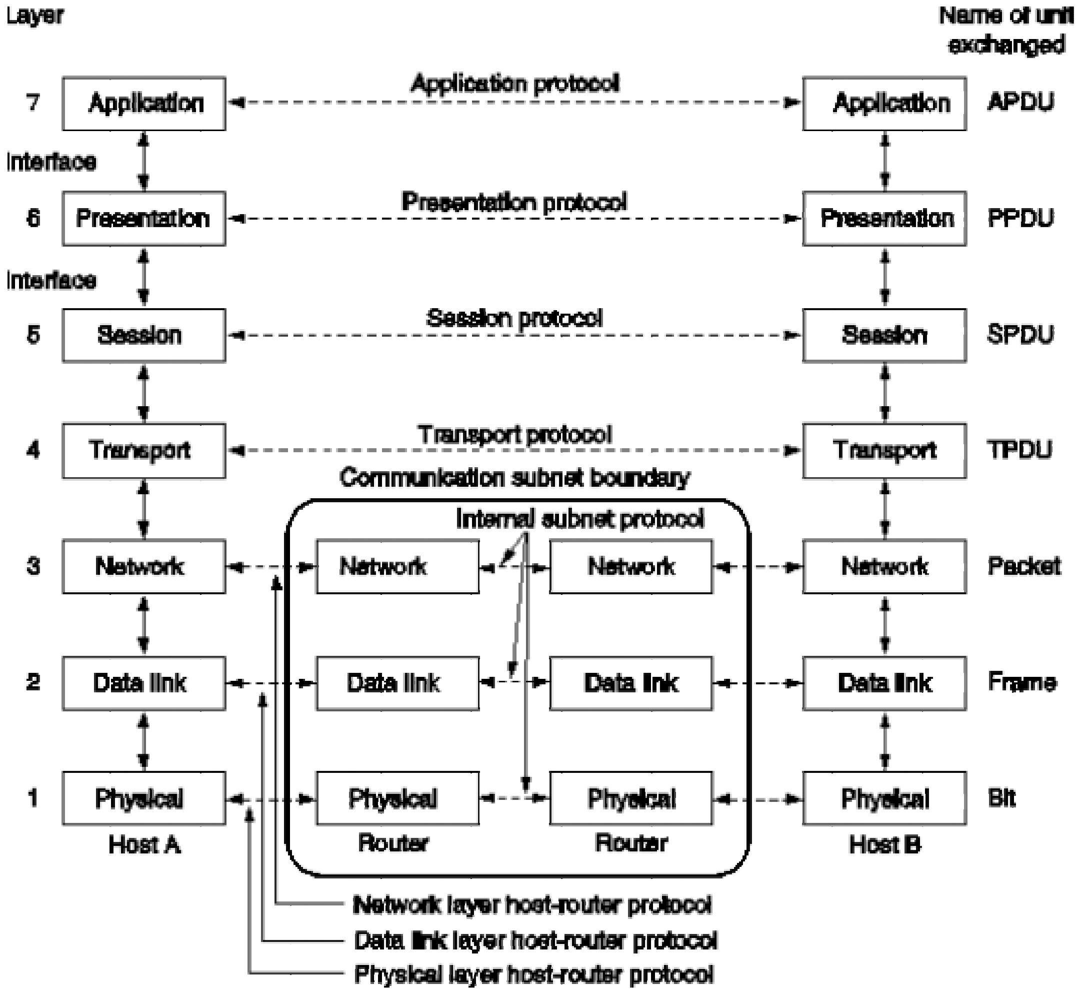


Fig. 1-16. The OSI reference model.

The OSI model is not a network architecture - it does not specify the exact services and protocols. It just tells what each layer should do. However, ISO has also produced standards for all the layers as a separate international standards.

1.4.2. The Physical Layer

The main task of the physical layer is to transmit raw bits over a communication channel.

Typical questions here are:

- how many volts should be used to represent 1 and 0,
- how many microseconds a bit lasts,
- whether the transmission may proceed simultaneously in both directions,
- how the initial connection is established and how it is turn down,
- how many pins the network connector has and what each pin is used for.

The design issues deal with mechanical, electrical, and procedural interfaces, and the physical transmission medium, which lies below the physical layer.

The user of the physical layer may be sure that the given stream of bits was encoded and transmitted. He cannot be sure that the data came to the destination without error. This issue is solved in higher layers.

1.4.3. 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. To accomplish this, the sender breaks the input data into data frames (typically a few hundred or a few thousand bytes), transmits the frames sequentially, and processes the acknowledgment frames sent back by the receiver.

The issues that the layer has to solve:

- to create and to recognize frame boundaries - typically by attaching special bit patterns to the beginning and end of the frame,
- to solve the problem caused by damaged, lost or duplicate frames (the data link layer may offer several different service classes to the network layer, each with different quality and price),
- to keep a fast transmitter from drowning a slow receiver in data,
- if the line is bi-directional, the acknowledgment frames compete for the use of the line with data frames.

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 (medium access sublayer) deals with the problem.

The user of the data link layer may be sure that his data were delivered without errors to the neighbor node. However, the layer is able to deliver the data just to the neighbor node.

1.4.4. The Network Layer

The main task of the network layer is to determine how data can be delivered from source to destination. That is, the network layer is concerned with controlling the operation of the subnet.

The issues that the layer has to solve:

- to implement the routing mechanism,
- to control congestions,
- to do accounting,
- to allow interconnection of heterogeneous networks.

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

The user of the network layer may be sure that his packet was delivered to the given destination. However, the delivery of the packets needs not to be in the order in which they were transmitted.

1.4.5. 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 them to the network layer, and ensure that the pieces all arrive correctly at the other end. All this must be done in a way that isolates the upper layers from the inevitable changes in the hardware technology.

The issues that the transport layer has to solve:

- to realize a transport connection by several network connections if the session layer requires a high throughput or multiplex several transport connections onto the same network connection if network connections are expensive,
- to provide different type of services for the session layer,
- to implement a kind of flow control.

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. In lower layers, the protocols are between each machine and its immediate neighbors.

The user of the transport layer may be sure that his message will be delivered to the destination regardless of the state of the network. He need not worry about the technical features of the network.

1.4.6. 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.

Some of these services are:

- *Dialog control* - session can allow traffic to go in both directions at the same time, or in only one direction at a time. If traffic can go only in one way at a time, the session layer can help to keep track of whose turn it is.
- *Token management* - for some protocols it is essential that both sides do not attempt the same operation at the same time. The session layer provides tokens that can be exchanged. Only the side holding the token may perform the critical action.
- *Synchronization* - by inserting checkpoints into the data stream the layer eliminates problems with potential crashes at long operations. After a crash, only the data transferred after the last checkpoint have to be repeated.

The user of the session layer is in similar position as the user of the transport layer but having larger possibilities.

1.4.7. The Presentation Layer

The presentation layer perform certain functions that are requested sufficiently often to warrant finding a general solution for them, rather than letting each user solve the problem. This layer is, unlike all the lower layers, 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. Different computers may use different ways of internal coding of characters or numbers. 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.

1.4.8. 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. If they have to be used for a work with a full screen editor, many problems arise from their incompatibility. One way to solve this problem is to define network virtual terminal and write editor for this terminal. 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. All the virtual terminal software is in the application layer.

Another application layer function is file transfer. It must handle different incompatibilities between file systems on different computers. Further facilities of the application layer are electronic mail, remote job entry, directory lookup ant others.

1.4.9. Data Transmission in the OSI Model

Figure 1-17 shows an example how data can be transmitted using OSI model.

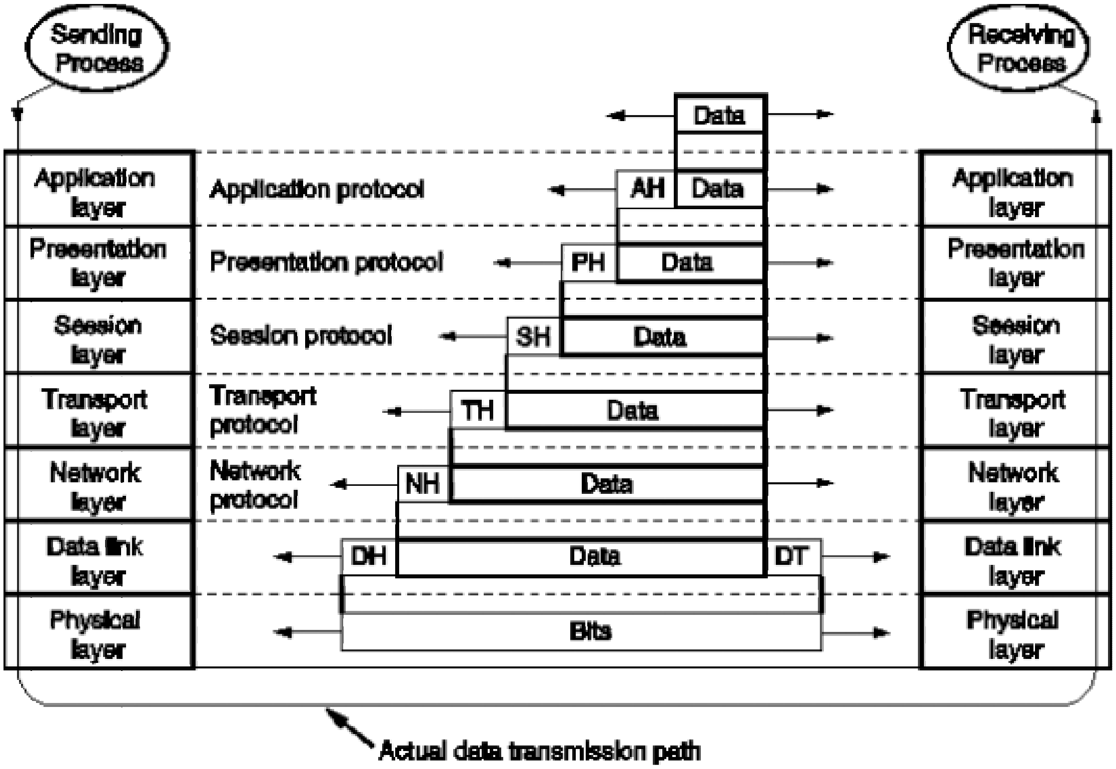


Fig. 1-17. An example of how the OSI model is used. Some of the headers may be null. (Source H.C. Folts. Used with permission.)

The key idea throughout is that although actual data transmission is vertical in Fig. 1-17, each layer is programmed as though it were horizontal. When the sending transport layer, for example, gets a message from the session layer, it attaches a transport header and sends it to the receiving transport layer. From its point of view, the fact that it must actually hand the message to the network layer on its own message is an unimportant technicality.

1.4.10. The TCP/IP Reference Model

TCP/IP reference model originates from the grandparent of all computer networks, the ARPANET and now is used in its successor, the worldwide Internet.

The name TCP/IP of the reference model is derived from two primary protocols of the corresponding network architecture.

1.4.11. The Internet Layer

The internet layer is the linchpin of the whole architecture. It is a connectionless internetwork layer forming a base for a packet-switching network. Its job is to permit hosts to inject packets into any network and have them travel independently to the destination. It works in analogy with the (snail) mail system. A person can drop a sequence of international letters into a mail box in one country, and with a little luck, most of them will be delivered to the correct address in the destination country.

The internet layer defines an official packet format and protocol called IP (Internet Protocol). The job of the internet layer is to deliver IP packets where they are supposed to go. TCP/IP internet layer is very similar in functionality to the OSI network layer (Fig. 1-18).

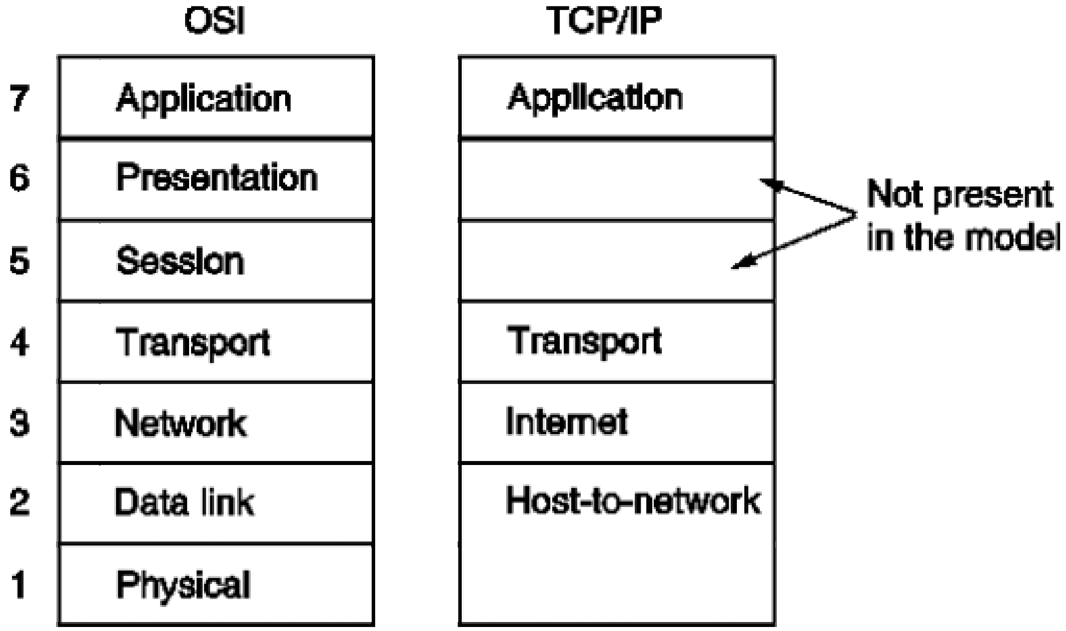


Fig. 1-18. The TCP/IP reference model.

1.4.12. The Transport Layer

The layer above the internet layer in the TCP/IP model is now usually called transport layer. It is designed to allow peer entities on the source and destination hosts to carry on a conversation, the same as in the OSI transport layer. Two end-to-end protocols have been defined here:

- *TCP* (Transmission Control Protocol) is a reliable connection-oriented protocol that allows a byte stream originating on one machine to be delivered without error on any other machine in the internet. It fragments the incoming byte stream into discrete messages and passes each one onto the internet layer. At the destination, the receiving TCP process reassembles the received messages into the output stream. TCP also handles flow control.
- *UDP* (User Datagram Protocol) is an unreliable, connectionless protocol for applications that do not want TCP's sequencing or flow control and wish to provide their own. It is also widely used for one/shot, client/server type request/reply queries and applications in which prompt delivery is more important than accurate delivery.

1.4.13. The Application Layer

The application layer is on the top of the transport layer. It contains all the higher level protocols. Some of them are:

- Virtual terminal (TELNET) - allows a user on one machine to log into a distant machine and work there.

- File transfer protocol (FTP) - provides a way to move data efficiently from one machine to another.
- Electronic mail (SMTP) - specialized protocol for electronic mail.
- Domain name service (DNS) - for mapping host names onto their network addresses.

1.4.14. The Host-to-Network Layer

Bellow the internet layer there is a great void. The TCP/IP reference model does not really say much about what happens here, except to point out that the host has to connect to the network using some protocol so it can send IP packet over it. This protocol is not defined and varies from host to host and network to network.

1.4.15. The ARPANET Story

The *ARPANET* is the grandparent of all computer networks, the Internet is its successor. The milestones of the ARPANET:

- In the mid 1960's, at the height of the Cold War, Department of Defense (DoD) wanted a command and control network that could survive a nuclear war. To solve this problem, DoD turned to its research arm Advanced Research Project Agency (ARPA).
- ARPA was created in response to the Soviet Union's launching Sputnik in 1957 and had the mission of advancing technology that might be useful to the military. It did its work by issuing grants and contracts to universities and companies whose ideas looked promising to it.
- ARPA decided that that the network the DoD needed should be a packet-switched network consisting of a subnet and host computers. The subnet would consist of minicomputers called IMPs (Interface Message Processors) connected by transmission lines. Each IMP would be connected to at least two other IMPs. At each IMP, there would be a host.
- ARPA put a tender for building the subnet and selected BBN, a consulting firm in Cambridge, Massachusetts for building the subnet and write the subnet software. The contract was signed in December 1968.
- BBN chose to use specially modified Honeywell DDP-316 minicomputers with 12K 16-bit words of memory as the IMPs. They did not have disks and were interconnected by 56 kbps lines leased from telephone companies.
- The software was split into two parts: subnet and host. The subnet software consisted of IMP end of the host-IMP connection, the IMP-IMP protocol, and a source IMP to destination IMP protocol. (Fig. 1-24).

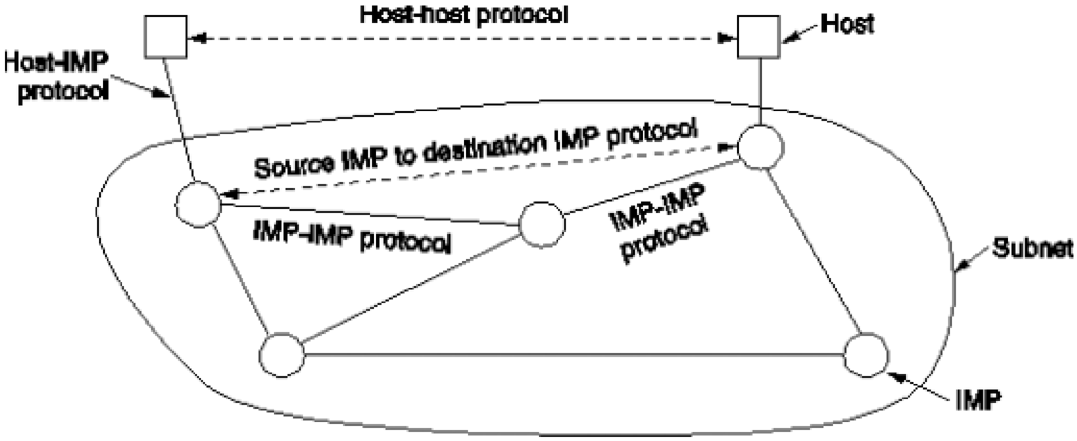


Fig. 1-24. The original ARPANET design.

- Host end of the host-IMP connection and host-host protocol as well as application software was written mostly by graduate students (BBN did not think it was their job).
- The experimental network with 4 nodes went on air in December 1969 and grew quickly (Fig. 1-25).

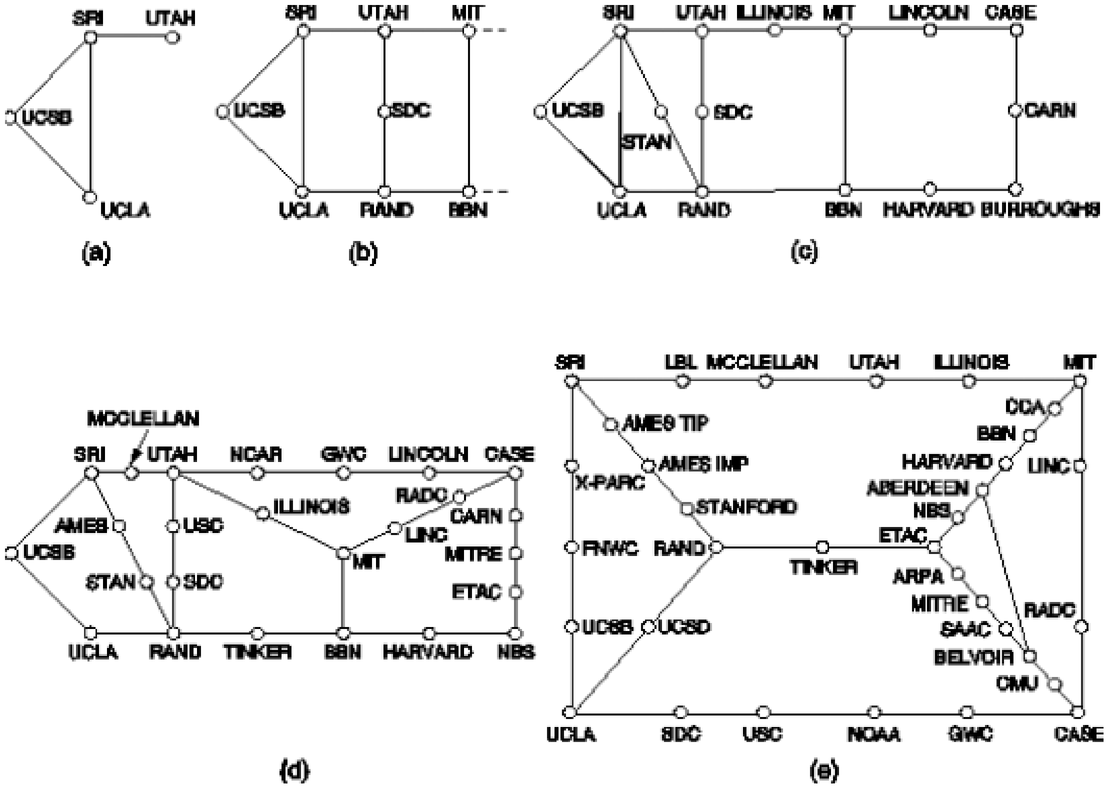


Fig. 1-25. Growth of the ARPANET. (a) Dec. 1969. (b) July 1970. (C) March 1971. (d) April 1972. (e) Sept. 1972.

- ARPA also funded research on satellite networks and mobile packet radio networks. In one famous demonstration a truck driving around California was connected with a computer in University College in London using packet radio network, ARPANET and satellite network.
- It turned out that ARPANET protocols were not suitable for running over multiple networks. This observation led to the invention of the TCP/IP model and protocols (Cerf and Kahn, 1974) specifically designed to handle communication over internetworks.
- University of California at Berkley integrated these new protocols to Berkley UNIX. The timing was perfect - many universities had just acquired new VAX computers with no networking software - they started to use Berkley software. With this software it was easy to connect to ARPANET.
- By 1983, the ARPANET was stable and successful, with over 200 IMPs and hundreds of hosts. At this point ARPA the military portion (about 160 IMPs) was separated into a separate subnet MILNET, with stringent gateways between MILNET and the remaining research subnet.
- During 1980s, additional networks were connected to ARPANET. DNS (Domain Naming System) was created to organize machines into domains and map host names onto IP addresses.
- By 1990, the ARPANET has been overtaken by newer networks that it itself has spawned, so it was shut down and dismantled, but it lives in the hearts and minds of network researchers everywhere. MILNET continues to operate.

1.4.16. A Comparison of the OSI and TCP Reference Models

The OSI and the TCP/IP reference models have much in common:

- they are based on the concept of a stack of independent protocols,
- they have roughly similar functionality of layers,
- the layers up and including transport layer provide an end-to-end network-independent transport service to processes wishing to communicate.

The two models also have many differences (in addition to different protocols).

Probably the biggest contribution of the OSI model is that it makes the clear distinction between its three central concepts that are services, interfaces, and protocols.

Each layer performs some services for the layer above it. The service definition tells what the layer does, not how entities above it access it or how the layer works.

A layer's interface tells the processes above it how to access it including the specification of the parameters and the expected results. But it, too, says nothing about how the layer works inside.

The peer protocols used in a layer are its own business. It can use any protocol as long as it provides the offered services.

These ideas fit with modern ideas about object-oriented programming where a layer can be understood to be an object with a set of operations that processes outside the object can invoke.

The TCP/IP model did not originally clearly distinguish between service, interface, and protocol. As a consequence, the protocol in the OSI model are better hidden than in the TCP/IP model and can be replaced relatively easily as the technology changes.

The OSI reference model was devised before the protocols were invented. The positive aspect of this was that the model was made quite general, not biased toward one particular set of protocols. The negative aspect was that the designers did not have much experience with the subject and did not have a good idea of which functionality to put into which layer (e.g. some new sublayers had to be hacked into the model).

With the TCP/IP the reverse was true: the protocols came first, and the model was just a description of the existing protocols. As a consequence, the model was not useful for describing other non-TCP/IP networks.

An obvious difference between the two models is the number of layers. Another difference is in the area of connectionless versus connection-oriented communication. The OSI model supports both types of communication in the network layer, but only connection-oriented communication in the transport layer. The TCP/IP model has only connectionless mode in the network layer but supports both modes in the transport layer. The connectionless choice is especially important for simple request-response protocols.

1.4.17. A Critique of the OSI Model and Protocols

At the end of 80s, it appeared that the OSI model were going to take over the world. This did not happen. The main reasons can be summarized as:

1. Bad timing.
2. Bad technology.
3. Bad implementation.
4. Bad politics.

1.4.18. Bad Timing

The time at which a standard is established is absolutely critical to its success (a theory of the apocalypse of the two elephants). The standard for a new subject has to be written between the two "elephants": the burst of research activities on the new subject and the burst of investments to the new subject. If it is written too early, before the research is finished, the subject may still be poorly understood, which leads to bad standard. If it is written too late, companies have already made investment and the standard is ignored. If the interval between the two elephants is very short, the people developing the standard may get crushed.

It appears that the standard OSI got crushed because of the use of TCP/IP protocols by research universities by the time OSI protocols appeared. At that time many vendors had already begun offering TCP/IP products and did not want to support a second protocol stack until they were forced to, so there were no initial offerings. With every company waiting for every other company to go first, no company went first and OSI never happened.

1.4.19. Bad Technology

The OSI model and the protocols are imperfect. Some layers are of little use or almost empty (the session, or the presentation layer), some are so full that subsequent work has split them into multiple sublayers, each with different functions (the data link, or the network layers). The real reason for 7 layers probably was that IBM had at the time when the OSI model was designed its proprietary seven-layered protocol called SNA (System Network Architecture).

The OSI model is extraordinarily complex. It is difficult to implement and inefficient in operation.

Perhaps the most serious criticism is that the model is dominated by a communications mentality.

1.4.20. Bad Implementation

Given the enormous complexity of the model and protocols, the initial implementations were huge, unwieldy, and slow. While the products got better in the course of time, the image stuck.

In contrast, the implementations of TCP/IP were good. People began to use them quickly which led to a large user community, which led to improvements, which led to an even large community and the spiral was upward.

1.4.21. Bad Politics

Many people, especially in academia, thought of TCP/IP as a part of UNIX, and UNIX in 1980s in academia was very popular.

OSI, on the other hand, was thought to be the creature of bureaucrats trying to shove a technically inferior standard down the throats of the poor researchers and programmers. It did not OSI help much.

But there are still a few organizations interested in OSI. Consequently, an effort has been made to update it, resulting in a (little) revised model published in 1994.

1.4.22. A Critique of the TCP/IP Reference model

The TCP/IP model and protocols have their problems to. The main of them are:

- the model does not clearly distinguish the concepts of service, interface, and protocols (it does not fit into good software engineering practice).
- TCP/IP model is not at all general and therefore it is poorly suited to describing any protocol stack other than TCP/IP.
- The host-to-network layer is not really a layer at all in the normal sense. It is an interface between the network and data link layers.
- The TCP/IP model does not distinguish, or even mention, the physical and data link layers.
- Although the IP and the TCP protocols were carefully thought out, and well implemented, many of the other protocols were ad hoc, produced by a couple of graduate students hacking away until they got tired. They were distributed free, widely used, deeply entrenched, and thus hard to replace. Some of them are a bit of embarrassment now (TELNET was designed for slow terminals, it knows nothing of graphical user interface and mice, but it is still widely used).

In summary, despite its problems, the OSI model (minus the session and presentation layers) has proven to be exceptionally useful for discussing computer networks. In contrast, the OSI protocols have not become popular. The reverse is true of TCP/IP: the model is practically nonexistent, but the protocols are widely used.

1.5. Example networks

Numerous network are currently operating around the world:

- public networks run by common carriers or PTTs,
- research networks,
- cooperative networks run by their users,
- commercial or corporate networks.

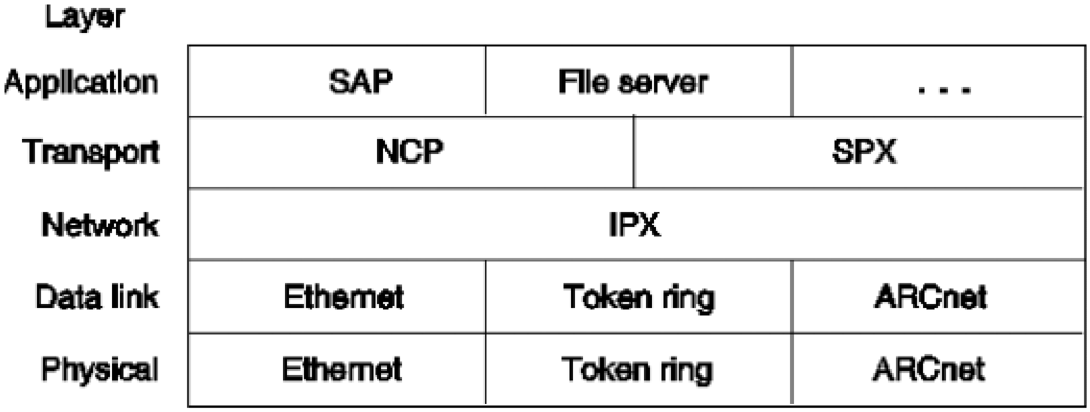
Networks differs in their:

- history,
- administration, facilities offered, technical design, user communities.

1.5.1. Novell NetWare

Novell NetWare is the most popular network system in the PC world. It was designed to be used by companies downsizing from a mainframe to a network of PCs. Novell NetWare is based on the client-server model.

NetWare uses a proprietary protocol stack (Fig. 1-22). It looks more like TCP/IP than like OSI.



<table>
<tr><th>Layer</th><th colspan="6"></th></tr>
<tr><td>Application</td><td colspan="2">SAP</td><td colspan="2">File server</td><td colspan="2">. . .</td></tr>
<tr><td>Transport</td><td colspan="3">NCP</td><td colspan="3">SPX</td></tr>
<tr><td>Network</td><td colspan="6">IPX</td></tr>
<tr><td>Data link</td><td colspan="2">Ethernet</td><td colspan="2">Token ring</td><td colspan="2">ARCnet</td></tr>
<tr><td>Physical</td><td colspan="2">Ethernet</td><td colspan="2">Token ring</td><td colspan="2">ARCnet</td></tr>
</table>

Fig. 1-22. The Novell Netware reference model.

The physical and data layers can be chosen from among various industry standards (Ethernet, IBM token ring, ARCnet).

The network layer runs an unreliable internetwork connectionless protocol called IPX, functionally similar to IP.

Above IPX comes a connection-oriented transport protocol called NCP (Network Core Protocol) providing various other services besides user data transport. A second protocol, SPX, is also available, but provides only transport. The session and presentation layers do not exist. Various application protocols are present in the application layer.

The IPX packet consists of the following fields:

- Checksum (2 bytes) - rarely used, since the underlying data link layer also provides a checksum.
- Packet length (2 bytes) - determines how long the entire packet is.
- Transport control (1 byte) - counts, how many networks the packet has traversed. When it exceeds a maximum, the packet is discarded.
- Packet type (1 byte) - used to mark various control packets.
- 2 address fields (12 bytes each) - each contains a 32-bit network number, a 48 bit machine number (the 802 LAN address), and 16 bit local address (socket) on that machine.
- Data.

The maximum size of the packet is determined by the underlying network.

About once a minute, each server broadcasts a packet giving its address and telling what services it offers (by using SAP - Service Advertising Protocol). The packets are collected by special agent processes running on the router machines. The agents use the information contained in them to construct databases of which servers are running where.

When a client machine is booted, the following procedures take place:

1. The client machine broadcasts a request asking where the nearest server is.
2. The agent on the local router machine looks in its database of servers and the best choice of server send back to the client.
3. The client establishes an NCP connection with the server. From this point on, the client can access the file system and other services using this connection.

1.5.2. NSFNET

NSF (the US National Science Foundation), seeing an enormous impact of the ARPANET, set up, by the late 1970s, a virtual network CSNET. It was centered around a single machine, supported dial-up lines, and had connections to the ARPANET and other networks. Using CSNET, academic researchers could call up and leave e-mail for other people to pick up later.

By 1984, NSF began designing a high-speed successor to the ARPANET for all university research groups. First, the supercomputer centers in San Diego, Boulder, Champaign, Pittsburgh, Ithaca, and Princeton were connected establishing the backbone of the network. Each supercomputer was given a little brother, consisting of an LSI-11 microcomputer called a fuzzball. The fuzzballs were connected with 56 kbps leased lines and formed the subnet, the same hardware technology as the ARPANET used. The software technology was different however: the fuzzballs spoke TCP/IP right from the start, making it the first TCP/IP WAN.

NSF also funded about 20 regional networks that connected to the backbone to allow users at thousands of universities, research labs, libraries, and museums to access any of the computers and to communicate with one another. The complete network, including the backbone and the regional networks, was called NSFNET. It was connected to the ARPANET through a link in the Carnegie-Mellon machine room (Fig. 1-26).

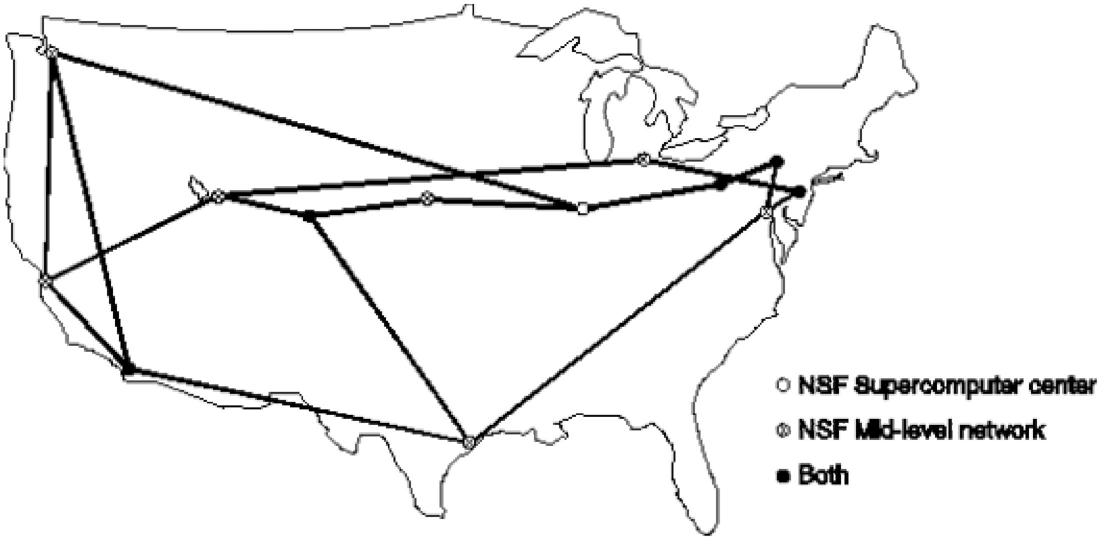


Fig. 1-26. The NSFNET backbone in 1988.

NSFNET was an instantaneous success and was overloaded from the word go. NSF immediately began planning its successor and the second version of the backbone was based on fiber optic channels at 448 kbps. In 1990, the second backbone was upgraded to 1.5 Mbps.

In 1990, a nonprofit corporation ANS (Advanced Networks and Services), initiated by NSF, took over the NSFNET and upgraded the 1.5 Mbps links to 45 Mbps to form ANSNET.

By 1995, the NSFNET backbone was no longer needed to interconnect NSF regional networks because numerous companies were running commercial IP networks. ANSNET was sold to America Online in 1995 and regional networks had to buy commercial IP services to interconnect.

To ease the transition, NSF awarded contracts to four different network operators to establish a NAP (Network Access Point). These NAP were established in San Francisco, Chicago, Washington and New York. Every network operator that wanted to provide backbone services to the NSF regional networks had to connect to all the NAPs. Consequently the network carriers were forced to compete for the regional networks business on the basis of service and price. The concept of single default backbone was replaced by a commercially driven competitive infrastructure.

In December 1991, the U.S. Congress passed a bill authorizing NREN, the National Research and Educational Network, the research successor to NSFNET, only running at gigabit speeds. The goal was a national network running at 3Gbps before the millennium. This network is to act as a prototype for the much-discussed information superhighway.

Other countries and regions are also building networks comparable to NSFNET. In Europe, EBONE is an IP backbone for research organizations and EuropaNET is a more commercially oriented network. Both connect numerous cities in Europe with 2 Mbps lines. Upgrades to 34 Mbps are in progress. Each country in Europe has one or more national networks, which are roughly comparable to the NSF regional networks.

1.5.3. The Internet

The number of networks, machines, and users connected to the ARPANET grew rapidly after TCP/IP became the only official protocol on Jan. 1, 1983. When NSFNET and ARPANET were interconnected, the grows became exponential. Connection were also made to networks in Canada, Europe, and the Pacific.

Sometime in the mid-1980s, people began viewing the collection of networks as an internet, and later as the Internet, although there was no official dedication with some politician breaking a bottle of champagne over a fuzzball.

Some facts about the growth of the Internet:

- In 1990, 3.000 networks, 200.000 computers.
- In 1992, the one millionth host was attached.
- By 1995, multiple backbones, hundreds of mid-level (regional) networks, tens of thousands of LANs, millions of hosts, and tens of millions of users. The size doubles approximately every year.

The glue that holds the Internet together is the TCP/IP reference model and the TCP/IP protocol stack.

A definition what does it mean to be on the Internet: a machine is on the Internet if it runs the TCP/IP protocol stack, has an IP address, and has the ability to send and receive IP packets to all the other machines on the Internet.

With exponential growth, the old informal way of running the Internet no longer works. In January 1992, the Internet Society was set up, to promote the use of the Internet and eventually take over managing it.

Main applications provided by the Internet:

- *e-mail*. This service has been available since the early days of the ARPANET and is enormously popular.
- *News*. Newsgroups are specialized forums in which users with a common interest can exchange messages.
- *Remote login*. Users on the Internet can log into any other machine on the Internet on which they have account.
- *File transfer*. Users can copy files from one machine on the Internet to another.

Up until early 1990s, the Internet was largely populated by researchers. One new application, WWW (World Wide Web), changed all that and brought millions of new nonacademic users to the net. This application was invented by CERN physicist Tim Berners-Lee.

1.5.4. Gigabit Testbeds

The Internet backbones operate at megabit speeds. The next step is gigabit networking. With each increase in the network bandwidth, new application become possible, so it is wit gigabit networks.

Gigabit networks provide better bandwidth than megabit networks, but not much better delay. For example, sending a 1 Kbit packet from New York to San Francisco at 1 Mbps takes 1 msec to pump the bits out and 20 msec for the transcontinental delay, for total of 21 msec. A 1 Gbps network can reduce this to 20.001 msec (the bits go out faster, the transcontinental delay remains the same, given by the speed of light in optical fiber 200.000 km/sec independent of the data rate. So the gigabits networks may only help for wide area applications where the bandwidth is what counts and are not helpful for those, where low delay is critical.

Two of the possible gigabit applications are telemedicine (the transfer of high quality images for diagnostic purposes) and virtual meetings (using some methods of virtual reality).

Starting in 1989, ARPA and NSF jointly agreed to finance a number of university-industry gigabits testbed.

1.6. Example Data Communication Services

Telephone companies and others have begun to offer networking services to any organization that wishes to subscribe. The subnet is owned by the network operator, providing communication service for the customers' hosts and terminals. Such a system is called a public network. It is analogous to, and often a part of, the public telephone system.

1.6.1. X.25 Networks

Many older network follow a standard called X.25 developed during the 1970s by CCITT to provide an interface between public packet-switched networks and their customers.

The physical layer protocol, called X.21, specifies the physical, electrical, and procedural interface between the host and the network.

The network layer protocol allows the user to establish virtual circuits and then send packets of up to 128 bytes on them. These packets are delivered reliably and in order. Most X.25 networks work at speeds up to 64 kbps. They are obsolete but still widespread.

X.25 is connection-oriented and supports two kinds of virtual circuits:

1. Switched virtual circuit is created when one computer sends a packet to the network asking to make a call to a remote computer. Once established, it can be used for sending packets.
2. Permanent virtual circuit is set up in advance by agreement between the customer and the carrier. It is always present and no call setup is required to use it. It is analogous to a leased line.

X.25 networks make possible to connect also ordinary (nonintelligent) terminals. It is realized by means of PADs (Packet Assembler Disassembler) whose function is described in a document known as X.3. A standard protocol between the terminal and PAD is called X.28, the protocol between the PAD and the network is called X.29.

1.6.2. Frame Relay

Frame relay can best be thought of as a virtual leased line. The customer leases a permanent virtual circuit between two points and can send frames (i.e. packets) of up to 1600 bytes between them.

The difference between an actual leased line and a virtual leased line is that with an actual one, the user can send traffic all day long at the maximum speed. With a virtual one, data burst may be sent at full speed, but the long-term average usage must be below a predetermined level. In return, the carrier charges much less for a virtual line than a physical one.

Frame relay provides a minimal service. For example, it is up to the user to discover that a frame is missing and to take the necessary action to recover.

1.6.3. Broadband ISDN and ATM

The telephone companies are aced with fundamental problem: multiple networks. Telephone and Telex use old circuit-switched networks. Each of the new data services as frame relay uses its own packet-switched network. DQDB (MAN) is different from these, and there is also the internal telephone call management network. Maintaining all these separate networks is a major headache, and there is another network, cable television, that the telephone companies do not control and would like to.

The solution of this problem is to invent a single new network for the future that will replace all the specialized networks with a single integrated network for all kinds of information transfer. This new network will have a huge data rate compared to all existing networks and services and will make it possible to offer a large variety of new services. This big project is now under way.

The new wide area service is called *B-ISDN* (Broadband Integrated Services Digital Networks). It will offer:

- video on demand,
- live television from many sources,
- multimedia electronic mail,
- CD-quality music,
- LAN interconnection,
- high-speed data transport for science and industry,
- many other services, all over the telephone line.

The underlying technology that makes B-ISDN possible is called ATM (Asynchronous Transfer Mode) because it is not synchronous (tied to a master clock).

A great deal of work has already been done on ATM and on B-ISDN system, although there is more ahead.

The basic idea behind *ATM* is to transmit all information in small, fixed-size packet called cells. The cells are 53 bytes long, of which 5 bytes are header and 48 bytes are data. ATM as a service is sometimes called cell relay.

ATM networks are connection-oriented.

ATM networks are organized like traditional WANs, with lines and switches. The intended speeds for ATM networks are 155 Mbps and 622 Mbps, with possible gigabit speeds later. The 155 Mbps speed was chosen because this is about what is needed to transmit high definition television. The exact choice of 155.52 Mbps was made for the compatibility with AT&T's SONET transmission system (the 622 Mbps are 4 155 Mbps channels).

It is worth pointing out that different organizations involved in ATM have different (financial) interests (the long-distance telephone carriers and PTTs vs. computer vendors). All these competing interests do not make the ongoing standardization process any easier, faster, or more coherent. Also, politics within the organization standardizing ATM (The ATM Forum) have considerable influence on where ATM is going.

1.6.4. The B-ISDN ATM Reference Model

Broadband ISDN using ATM has its own reference model (Fig. 1-30). It consists of three layers, plus whatever the users want to put on top of that. The three layers are:

- Physical layer. It deals with the issues of the physical medium. ATM cells may be sent on a wire or fiber by themselves, but they may be also be packaged inside the data of other carrier systems. In other words, ATM has been designed to be independent of the transmission medium.
- ATM layer. It deals with cells and cell transport. It defines the layout of a cell. It also deals with establishment and release of virtual circuits. Congestion control is also located here.
- AAL (ATM Adaptation Layer). It allows users to send packets larger than a cell. The ATM layer interface segments these packets, transmits the cells individually, and reassembles them at the other end.

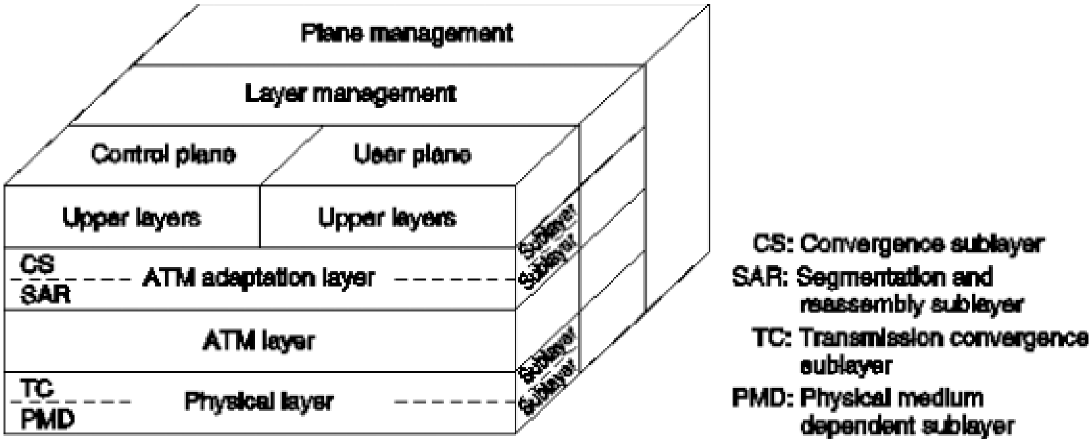
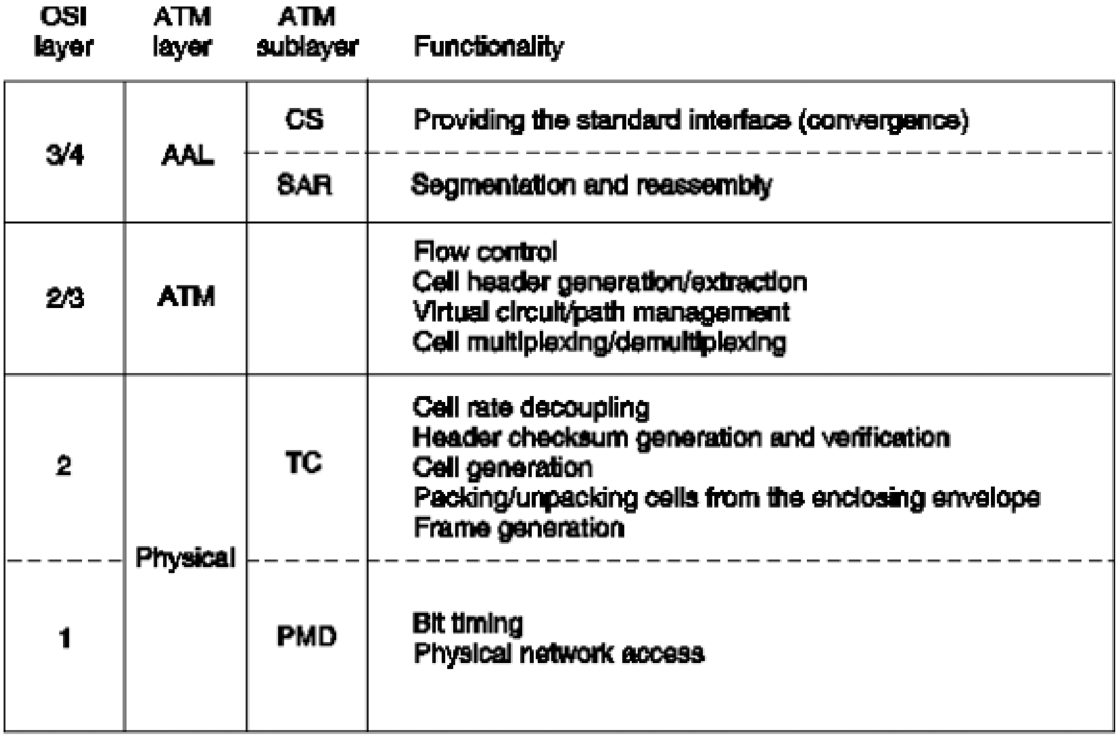


Fig. 1-30. The B-ISDN ATM reference model.

ATM model is defined as being three-dimensional. The user plane deals with data transport, flow control, error correction, and other user functions. The control plane is concerned with connection management. The layer and plane management functions relate to resource management and interlayer coordination.

The physical and AAl layers are each divided into two sublayers (Fig. 1-31):



OSI layer	ATM layer	ATM sublayer	Functionality
3/4	AAL	CS	Providing the standard interface (convergence)
		SAR	Segmentation and reassembly
2/3	ATM		Flow control Cell header generation/extraction Virtual circuit/path management Cell multiplexing/demultiplexing
2	Physical	TC	Cell rate decoupling Header checksum generation and verification Cell generation Packing/unpacking cells from the enclosing envelope Frame generation
1		PMD	Bit timing Physical network access

Fig. 1-31. The ATM layers and sublayers, and their functions.

- the bottom sublayer does the work,
- the top (convergence) sublayer provides the proper interface to the layer above it.

The PMD (Physical Medium Dependent) sublayer interfaces to the actual cable. It moves the bits on and off. For different carriers and cables, this layer will be different.

The TC (Transmission Convergence) sublayer sends the transmitted cells as a string to the PMD sublayer. At the other end, the TC converts a pure incoming bit stream from the PMD sublayer into a cell stream for the ATM layer (the task of data link layer of the OSI model).

ATM layer is a mixture of the OSI data link and network layers.

The SAR (Segmentation And Reassembly) sublayer breaks packets up into cells on the transmission side and puts them back together again at the destination.

The CS (Convergence Sublayer) makes it possible to have ATM systems offer different kinds of services to different applications (e.g. file transfer and video on demand have different requirements concerning error handling, timing , etc.).

1.6.5. Perspective on ATM

Some observations that will influence the future development on ATM:

- ATM is a project invented by the telephone industry, many computer vendors joined with the telephone companies to set up ATM Forum, that will guide the future of ATM.
- ATM is basically high-speed packet-switching, a technology the telephone companies have little experience with. They do have massive investment in a different technology (circuit switching). The transition will not happen quickly - it is a revolutionary change.
- Who will pay for the necessary replacement of the existing telephone system? How much will consumers be willing to pay to get a movie on demand electronically?

- Where the advanced services will be provided is crucial. If they are provided by the network, the telephone companies will profit from them. If they are provided by computers attached to the network, the manufacturer and operator of these devices make the profit.

2. The Physical Layer

The physical layer is the lowest layer in almost all reference models of computer networks.

2.1. The Theoretical Basis for Data Communication

Information is transmitted on wires by varying some physical property such as voltage or current. Let f(t) be a function of time representing the value of this voltage or current modeling the behavior of the signal.

2.1.1. Fourier Analysis

Any reasonable behaved periodic function, g(t), can be expressed in the form of Fourier series

$$g(t) = \frac{1}{2}c + \sum a_n \sin(2\pi nft) + \sum b_n \cos(2\pi nft)$$

where f = 1/T is the fundamental frequency and an and bn are the sine and cosine amplitudes of the n-th harmonics. The values of c, an, and bn can be expressed by the following equations:

$$c = \frac{2}{T}\int_0^T g(t)dt \qquad a_n = \frac{2}{T}\int_0^T g(t)\sin(2\pi nft)dt \qquad b_n = \frac{2}{T}\int_0^T g(t)\cos(2\pi nft)dt$$

A data signal that has a finite duration can be handled as a periodical function imagining that it repeats the entire pattern over and over.

2.1.2. Bandwidth-Limited Signals

Consider an example - the transmission of the ASCII character b encoded in an 8-bit byte as 01100010. The voltage output of the transmitting computer is shown in Fig. 2-1(a). The Fourier analysis of this signal yields the coefficients:

$$a_n = \frac{1}{\pi n}\left[\cos(\frac{\pi n}{4}) - \cos(\frac{3\pi n}{4}) + \cos(\frac{6\pi n}{4}) - \cos(\frac{7\pi n}{4})\right]$$

$$b_n = \frac{1}{\pi n}\left[-\sin(\frac{\pi n}{4}) + \sin(\frac{3\pi n}{4}) - \sin(\frac{6\pi n}{4}) + \sin(\frac{7\pi n}{4})\right]$$

$$c = \frac{3}{8}$$

The values $a_n^2 + b_n^2$ are of interest because they are proportional to the energy at the corresponding frequency (Fig. 2-1(a)).

No transmission facility can transmit signal without loosing some power in the process of transmission. All transmission facilities diminish different Fourier components by different amount, thus introducing distortion. If all Fourier components were equally diminished, the resulting signal would be reduced in amplitude but not distorted, i.e., it would have the same nice squared-off shape as in Fig. 2-1. Usually, the amplitudes are transmitted undiminished from 0 up to some frequency fc (measured in cycles/sec or Hertz (Hz)) with all frequencies above this cutoff frequency strongly attenuated (as a consequence of a physical property of transmission medium or intentionally introduced by filter).

Fig. 2-1 shows how the signal of Fig. 2-1(a) would look if the bandwidth were so low that only the lowest frequencies were transmitted.

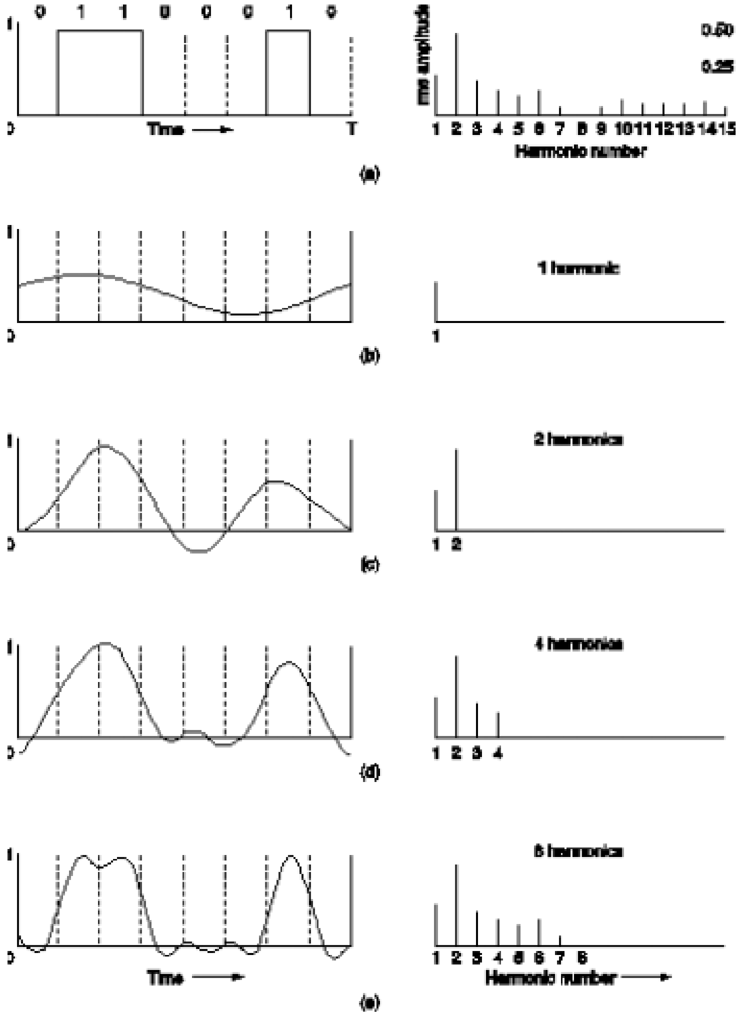


Fig. 2-1. (a) A binary signal and its root-mean square Fourier amplitudes.
(b)-(e) Successive approximations to the original signal.

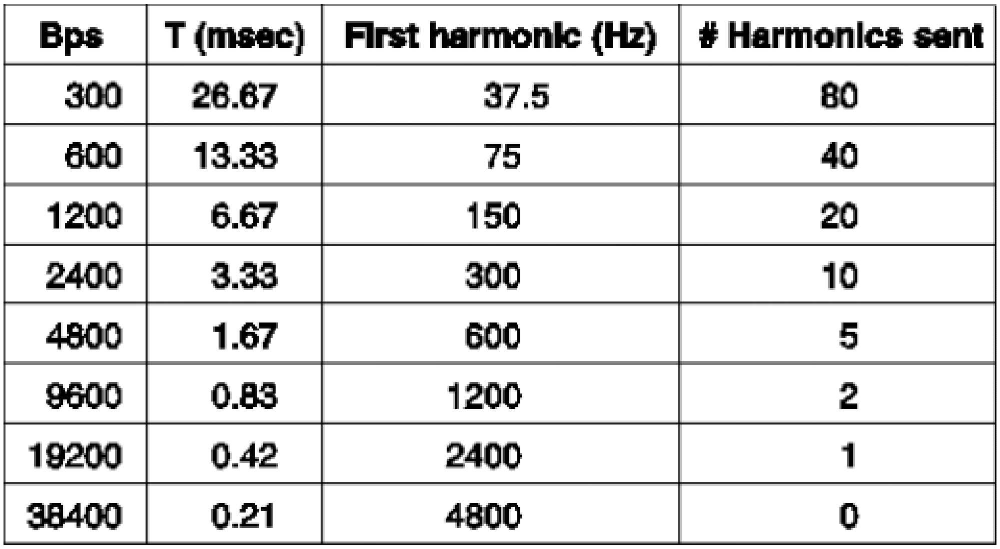
The time T required to transmit a character depends on:

- the encoding method,
- the signaling speed (the number of times per second that the signal changes its value).

The number of changes per second is measured in baud. A b baud line does not necessarily transmits b bits/sec since each signal might convey several bits. If the voltage 0,1,2,...,7 were used, each signal value could be used to convey 3 bits, so the bit rate would be three times the baud rate. In our example, only 0s and 1s are being used as a signal levels, so the bit rate is equal to baud rate.

If a bit rate is b bits/sec, the time to send an 8 bits character is 8/b. The frequency of character transmission is b/8. If we have a channel with a cutoff frequency f, the number of the highest harmonic passed through the channel is f/(b/8).

Example: An ordinary telephone line (called often voice-grade line), has an artificially introduced cutoff frequency near 3000 Hz. For some data rates, the number of the highest harmonics passed through the line are shown in Fig. 2-2.



Bps	T (msec)	First harmonic (Hz)	# Harmonics sent
300	26.67	37.5	80
600	13.33	75	40
1200	6.67	150	20
2400	3.33	300	10
4800	1.67	600	5
9600	0.83	1200	2
19200	0.42	2400	1
38400	0.21	4800	0

Fig. 2-2. Relation between data rate and harmonics.

Sophisticated coding schemes that use several voltage levels do exist and can achieve higher data rates than 38.4 kbps.

2.1.3. The Maximum Data Rate of a Channel

In 1924 H.Nyquist derived an equation expressing the maximum data rate for a finite bandwidth noiseless channel.

Nyquist proved that if an arbitrary signal has been run through a low-pass filter of bandwidth H, the filtered signal can be completely reconstructed by making only 2H exact samples per second. Sampling the line faster than 2H times per second is pointless because the higher frequency components that such sampling could recover have already been filtered out. If the signal consists of V discrete levels, Nyquist theorem states:

$$\text{maximum data rate} = 2H \log_2 V \text{ bits/sec}$$

For example, a noiseless 3-kHz channel cannot transmit binary (i.e. two-level) signal at rate exceeding 6000 bps.

If we consider the presence of noise, the situation is worse. If we denote the signal power by S and the noise power by N, as the measure of the signal-to-noise ratio the quantity 10 $\log_{10}$ S/N given in decibels (dB) is taken.

In 1948, Claude Shannon extended Nyquist’s work as follows: the maximum data rate of a noisy channel whose bandwidth is H Hz, and whose signal-to-noise ratio is S/N, is given by

$$\text{maximum number of bits/sec} = H \log_2 (1 + S/N)$$

For example, a channel of 3000 Hz bandwidth, and a signal to thermal noise ratio of 30 dB (S/N = 1000), can never transmit much more than 30000 bps, no matter how many or few signal levels are used. Shannon's result was derived using information-theory arguments and applies to any channel subject to Gaussian (thermal) noise.

2.2. Transmission Media

For the transmission of bit stream from one machine to another, various physical media can be used. They differ in terms of:

- bandwidth,
- delay,
- cost,
- easy of installation and maintenance.

Media can be divided into:

- *guided media* - copper wire, fiber optics,
- *unguided media* - radio, laser through the air.

2.2.1. Magnetic media

One of the most common ways to transport data from one computer to another is to write them onto magnetic tapes or floppy disks, physically transport the tapes or disks to the destination machine and read them back in again.

Example: Industry standard 8 mm video tape can hold 7 gigabytes. A box of 50 x 50 x 50 cm can hold about 1000 of these tapes for a total capacity 7000 GB. It can be delivered in 24 hours anywhere in the US. Effective bandwidth is 56000 gigabits/86400 sec = 648 Mbps (better than high-speed version of ATM (622 Mbps)). Estimated cost: 10 cents/gigabyte which is unbeatable. The disadvantage of this kind of the transmission is definitely big delay.

2.2.2. Twisted pairs

Twisted pair is the oldest and still most common transmission medium. It consists of two insulated copper wires, typically about 1 mm thick. The wires are twisted together to reduce electrical interference from similar pairs close by (two parallel wires constitute a simple antenna, a twisted pair does not).

The most common application of the twisted pair is the telephone system. Twisted pairs can run several km without amplification, but for longer distances repeaters are needed.

Twisted pairs can be used for either analog or digital transmission. The bandwidth depends on the thickness of the wire and the distance traveled (several mbps for a few km can be achieved).

Twisted pair cabling comes in several varieties, two of which are important for computer networks:

- Category 3 twisted pairs - gently twisted, 4 pairs typically grouped together in a plastic sheath.
- Category 5 twisted pairs - introduced in 1988. More twists per cm than category 3 and teflon insulation, which results in less crosstalk and better quality signal over longer distances.

2.2.3. Baseband Coaxial Cable

Coaxial cable (frequently called "coax") is another common transmission medium. It has better shielding than twisted pairs, so it can span longer distances at higher speeds.

Two kinds of coaxial cables are widely used:

- 50-ohm - used for digital transmissions,
- 75-ohms - used for analog transmissions.

(The distinction is based on historical rather than technical factors.)

A cutaway view of a coaxial cable is shown in Fig. 2-3. The bandwidth depends on the cable length. For 1 km cables, a data rate 1 - 2 Gbps is feasible. Longer cables enable only lower data rates or require periodic amplifiers.

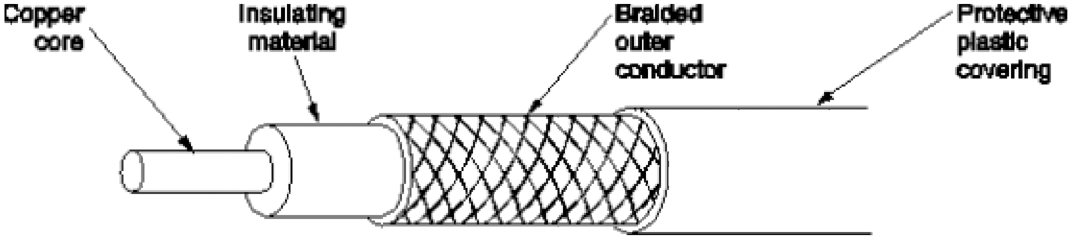


Fig. 2-3. A coaxial cable.

Coaxial cables used to be widely used within the telephone system - now they are largely replaced by fiber optics on long-haul routes (1000 km of fiber installed every day in the US).

2.2.4. Broadband Coaxial Cable

75-ohm coaxial cable is used on standard cable television. It is called broadband.

In the telephone world, "broadband" refers to anything wider than 4 kHz. In the computer networking world, "broadband cable" means any cable network using analog transmission (the analog signaling consists of varying voltage with time to represent an information stream).

The cables in broadband networks can be used often up to 450 MHz and can run for nearly 100 km due to the analog signaling which is much less critical than digital signaling. To transmit digital signal on an analog network, outgoing bit stream must be converted to an analog signal and the incoming analog signal to a bit stream. 1 bps may occupy roughly 1 Hz of the bandwidth. At higher frequencies, many bits per Hz are possible using advanced modulation techniques.

Broadband systems are divided up into multiple channels, frequently the 6-MHz channels used for television broadcasting. Each channel can be used for analog TV, CD quality audio (1.4 Mbps) or a digital bit stream at, say, 3 Mbps. Television and data can be mixed on the cable.

Amplifiers in broadcast systems can only transmit signal in one direction. When the cabling is used for connecting computers, so called dual cable systems and single cable systems have been developed (Fig. 2-4).

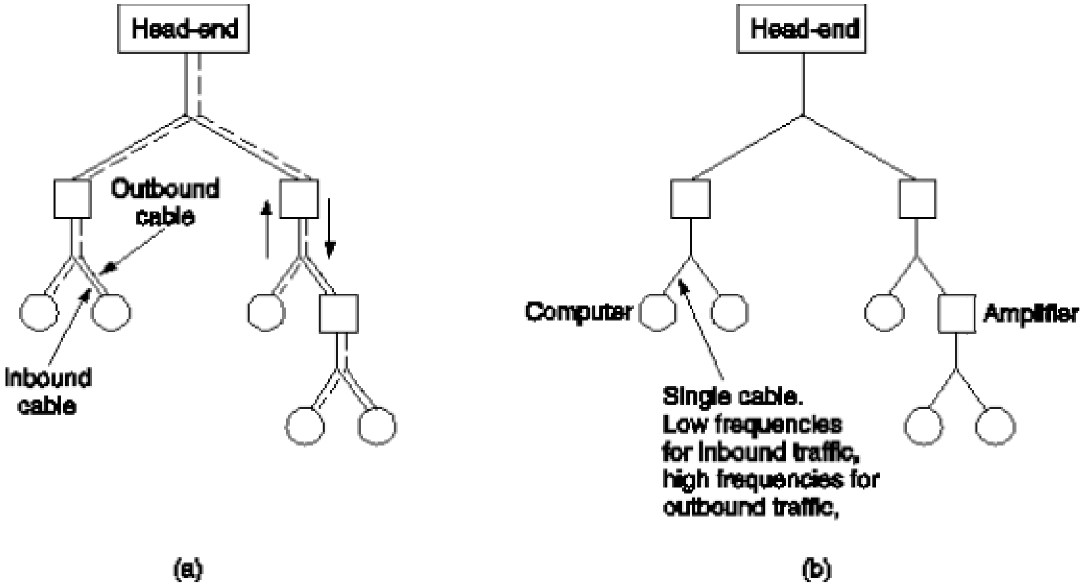

Head-end
Outbound cable
Inbound cable
(a)
Head-end
Computer
Amplifier
Single cable. Low frequencies for inbound traffic, high frequencies for outbound traffic,
(b)

Fig. 2-4. Broadband networks. (a) Dual cable. (b) Single cable.

2.2.5. Fiber Optics

In race between computing and communication, communication won (improvement factor 10 vs. 100 per decade during the last two decades) due to using fibre optics in communication.

An optical transmission system has three components:

- the light source - a pulse of light indicates a 1 bit and the absence of light indicates a 0 bit,
- the transmission medium - ultra-thin fiber of glass,
- the detector - generates an electrical pulse when light falls on it.

By attaching a light source to one end of an optical fiber and a detector to the other, we get a unidirectional data transmission system.

The work of this transmission system is based on the refraction of the light ray at the silica/air boundary (Fig. 2-5).

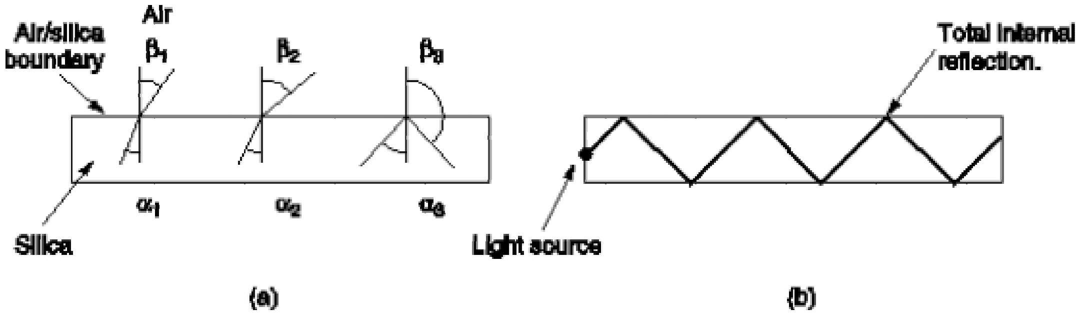


Fig. 2-5. (a) Three examples of a light ray from inside a silica fiber impinging on the air/silica boundary at different angles. (b) Light trapped by total internal reflection.

Since any light ray incident on the boundary above the critical angle will be reflected internally, many different rays will be bouncing around at different angles. Each ray is said to have a different mode, so a fiber having this property is called a multimode fiber.

If the fiber's diameter is reduced to a few wavelengths of light, the fibre acts like a wave guide and the light can only propagate in a straight line, without bouncing, yielding a single mode fiber.

Single mode fibers are more expensive but can be used for longer distances (typically several Gbps for 30 km).

2.2.6. Transmission of Light through Fiber

The glass used for modern optical fibers is so transparent that if the ocean were full of it instead of water, the seabed would be visible from the surface.

The attenuation of light through glass depends on the wavelength (Fig. 2-6). It is expressed in decibels given by the formula:

Attenuation in decibels = 10 log$_{10}$ transmitted power/received power

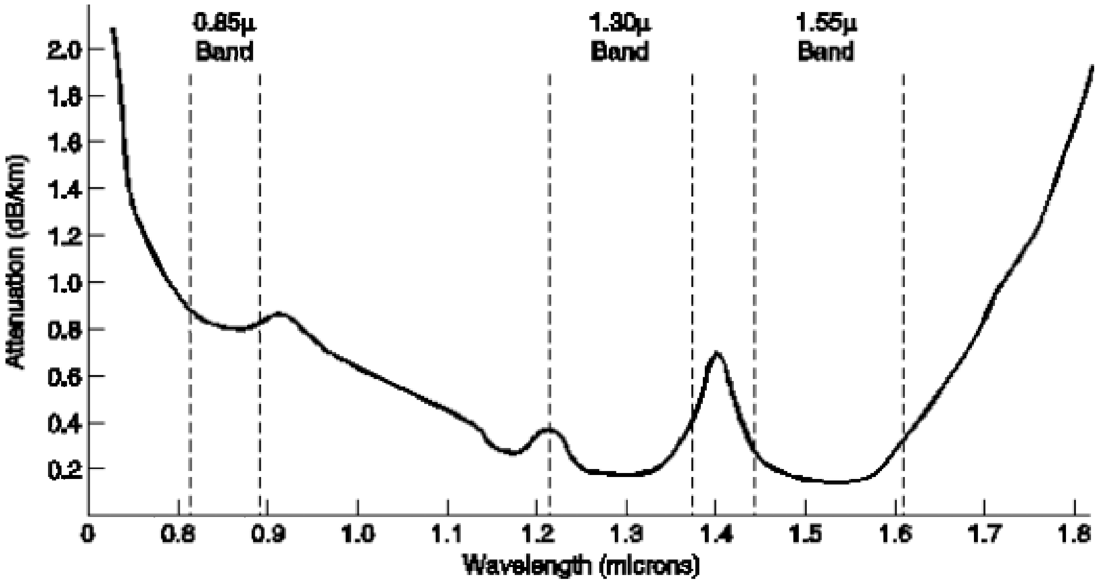


Fig. 2-6. Attenuation of light through fiber in the infrared region.

Wavelength 0.85, 1.30, and 1.55 microns (micro meters) are used for communication (Fig. 2-6). 0.85 has higher attenuation but it has a nice property that, at that wavelength, the lasers and electronics can be made from the same material. The bands for all three wavelengths are 25000 - 30000 GHz.

Visible light has slightly shorter wavelength (0.4 - 0.7 microns).

Light pulses sent down a fiber spread out in length as they propagate. This is called dispersion. The amount of dispersion is wavelength dependent. It was discovered (so far it works just in lab conditions) that when pulses have special shape (called solitons) all the dispersion effects cancel.

2.2.7. Fiber Cables

Fiber optics cables are similar to coax, except without the braid (Fig. 2-7). In multimode fibers, the core is typically 50 microns in diameter, in single mode fibers the core is 8 - 10 microns. The cladding has a lower index of refraction than the core to keep all the light in the core.

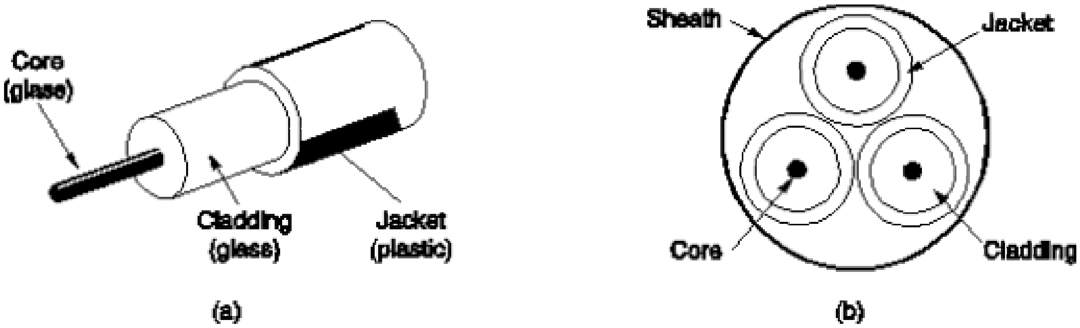


Fig. 2-7. (a) Side view of a single fiber. (b) End view of a sheath with three fibers.

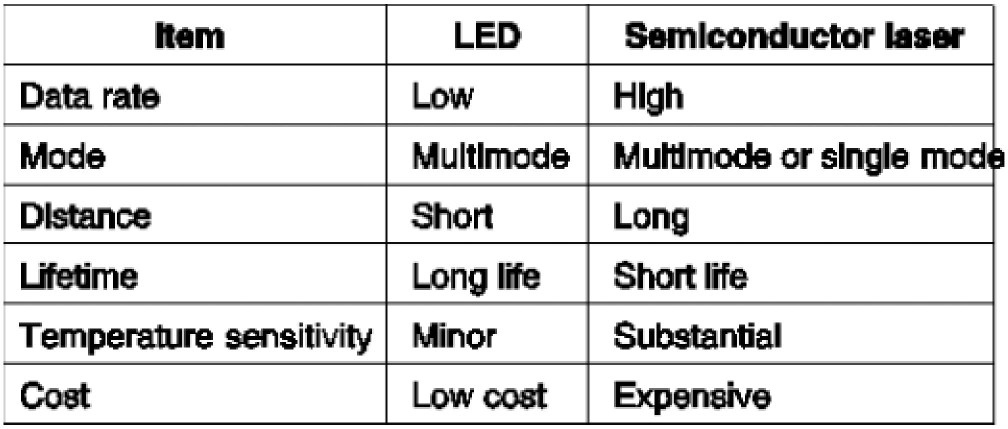
Fibers can be connected in three different ways:

- terminating in connectors and plugged into fiber sockets,
- spliced mechanically by a clamp,
- fused to form a solid connection.

Two kinds of light sources can be used to do the signaling:

- LEDs,
- semiconductor lasers.

The properties of the both sources are shown in Fig. 2-8.



Item	LED	Semiconductor laser
Data rate	Low	High
Mode	Multimode	Multimode or single mode
Distance	Short	Long
Lifetime	Long life	Short life
Temperature sensitivity	Minor	Substantial
Cost	Low cost	Expensive

Fig. 2-8. A comparison of semiconductor diodes and LEDs as light sources.

The receiving end of an optical fiber consists of a photo diode. The typical response time of a photodiode is 1 nsec which limits data rates to about 1 Gbps.

2.2.8. Fiber Optics Networks

Fiber optics can be used for LANs as well as for long-haul transmission. Tapping onto it is more complex than connecting to a copper wire. One way around the problem is shown in Fig. 2-9.

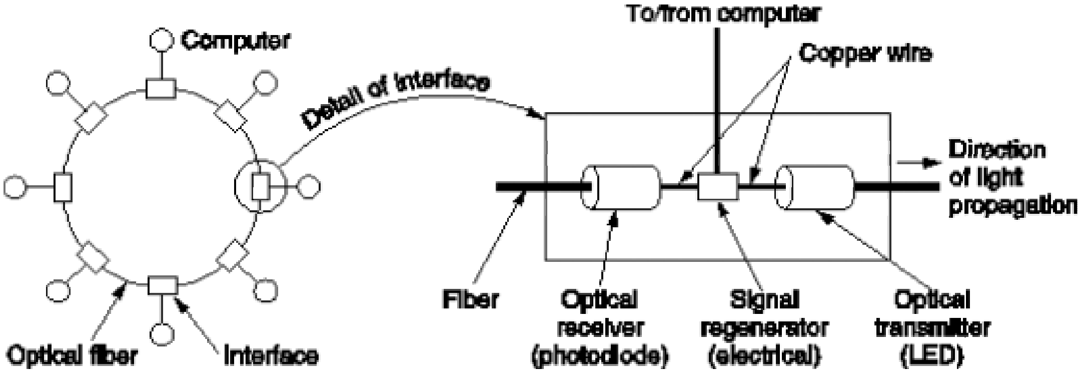

To/from computer
Computer
Copper wire
Detail of interface
Direction of light propagation
Fiber
Optical receiver (photodiode)
Signal regenerator (electrical)
Optical transmitter (LED)
Optical fiber
Interface

Fig. 2-9. A fiber optic ring with active repeaters.

Another solution is displayed in Fig. 2-10.

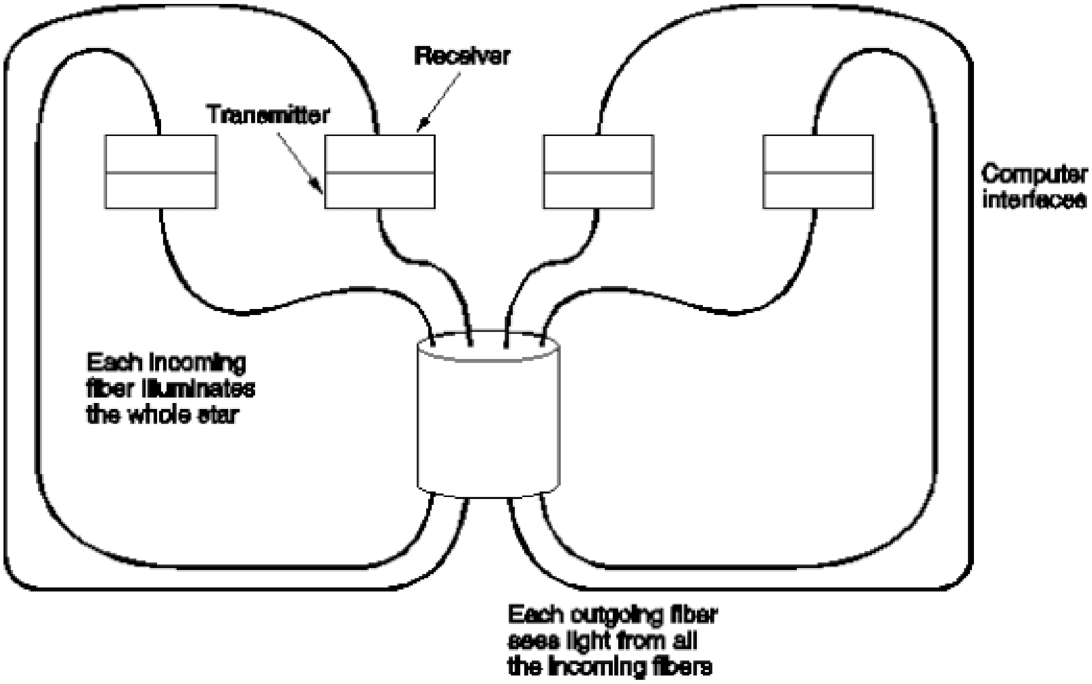

Receiver
Transmitter
Computer interfaces
Each incoming fiber illuminates the whole star
Each outgoing fiber sees light from all the incoming fibers

Fig. 2-10. A passive star connection in a fiber optics network.

2.2.9. Comparison of Fiber Optics and Copper Wire

Advantages of fibers:

- much higher bandwidth,
- low attenuation (30 km distance of repeaters vs. 5 km for copper),
- noise-resistance,
- not affected by corrosive chemicals,
- much lighter than copper - easier installation and maintenance,
- difficult to tap - higher security.

Disadvantages of fiber:

- unfamiliar technology so far,
- unidirectional communication,
- more expensive interfaces than electrical ones.

Nevertheless, the future of all fixed data communication for more than a few meters is clearly with fiber.

2.3. Wireless Transmission

Modern wireless digital communication began in the Hawaiian Islands, where large chunks of Pacific Ocean separated the users and the telephone system was inadequate.

2.3.1. The Electromagnetic Spectrum

When electrons move, they create electromagnetic waves that can propagate through free space. The number of oscillation per second of an electromagnetic wave is called its frequency, f, and measured in hertz (Hz). The distance of two consecutive maxima is called wavelength and universally designated by l (lambda).

By attaching an antenna of the appropriate size to an electrical circuit, the electromagnetic waves can be broadcasted efficiently and received by a receiver some distance away. All wireless communication is based on this principle.

In vacuum, all electromagnetic waves travel at the same speed, usually called the speed of light, c, approximately 3 x 10^8 m/sec. In copper or fiber the speed slows to about 2/3 of this value and becomes slightly frequency dependent.

The fundamental relation between f, l, and c (in vacuum) is

$$lf = c$$

For example: 1-MHz waves are about 300 m long and 1-cm waves have a frequency of 30 GHz.

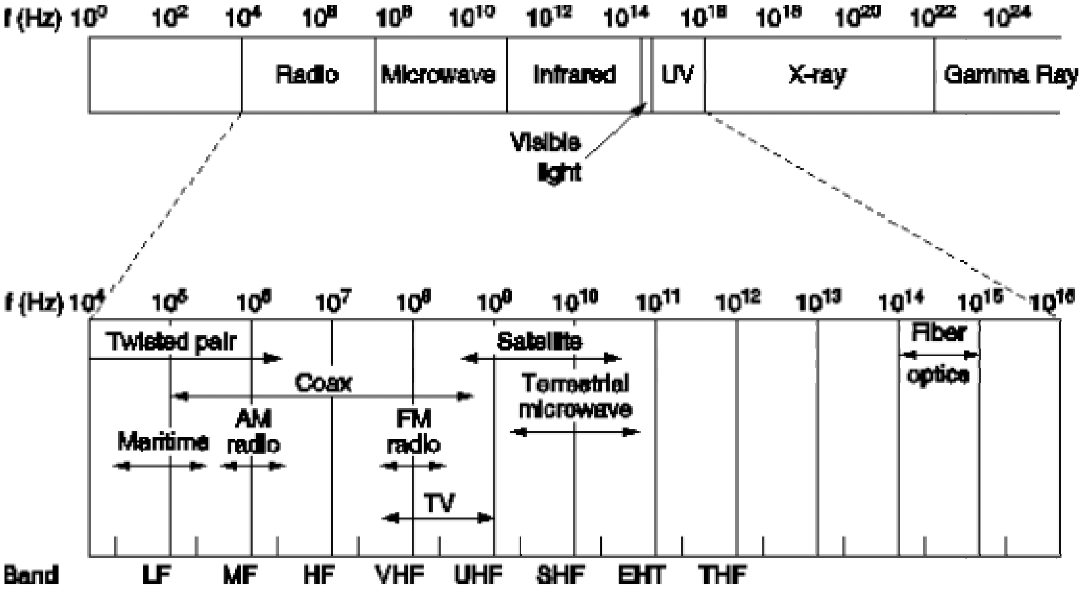


Fig. 2-11. The electromagnetic spectrum and its uses for communication.

The electromagnetic spectrum is shown in Fig. 2-11. The radio, microwave, infrared, and visible light portions of the spectrum can all be used for transmitting information by modulating the amplitude, frequency, or phase of the wave. Ultraviolet light, X-rays, and gamma rays would be even better, due to their higher frequencies, but they are hard to produce and modulate, do not propagate well through buildings, and are dangerous to living things.

LF, MF, ... are official ITU (International Telecommunication Union) names and are based on wavelengths.

The amount of information that an electromagnetic wave can carry is related to its bandwidth. With current technology, it is possible to encode a few bits per Hertz at low frequencies, but often as many as 40 under certain conditions at high frequencies, so a cable with 500 MHz bandwidth can carry several gigabits/sec.

There are national and international agreement about who gets to use which frequencies. World-wide, it is an agency of ITU-R (WARC), in US the work is done by FCC (Federal Communication Commission).

Most transmissions use a narrow frequency band ((f/f<<1)to get the best reception (many watts/Hz). However, there are some exception from this rule (i.e. spread spectrum popular in military communications).

2.3.2. Radio Transmission

Radio Waves are easy to generate, can travel long distances, and penetrate building easily, so they are widely used for communications, both indoors and outdoors. They are also omnidirectional, so the transmitter and receiver do not have to be aligned physically. This feature is sometimes good, but sometimes bad.

The properties of radio waves are frequency dependent. At low frequencies they pass through obstacles well, but the power falls off sharply with distance from the source. At high frequencies, radio waves tend to travel in straight lines and bounce off obstacles. They are also absorbed by rain. At all frequencies, they are subject to interference from motors and other electrical equipment.

Due to radio's ability to travel long distances, interference between users is a problem. For this reasons, all governments license the use user of radio transmitters.

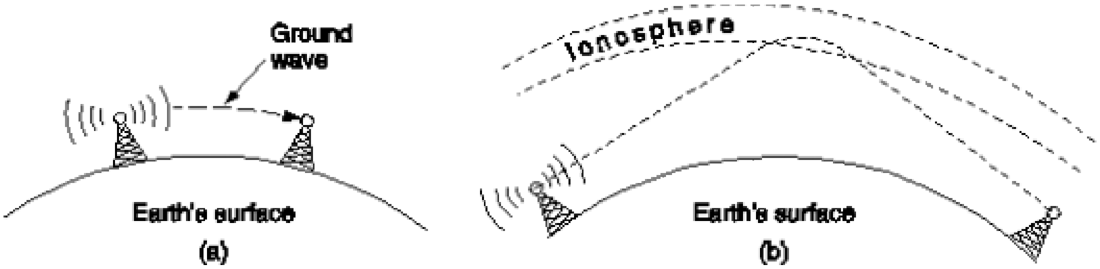


Fig. 2-12. (a) In the VLF, VF, and MF bands, radio waves follow the curvature of the earth. (b) In the HF they bounce off the ionosphere.

In the VLF, LF, and MF bands, radio waves follow the ground (Fig. 2-12(a)) and can be detected for about 1000 km at the lower frequencies, less at the higher ones. The main problem with using these bands for data communication is relatively low bandwidth they offer.

In the HF and VHF bands, the ground waves tend to be absorbed by the earth, but if they reach the ionosphere (a layer of charged particles circling the earth at a height of 100 to 500 km) are refracted (Fig. 2-12(b)) by it and sent back to earth. Amateur radio operators use these bands to talk long distance.

2.3.3. Microwave Transmission

Above 100 MHz, the waves travel in straight lines and can therefore be narrowly focused. Concentrating all the energy into a small beam using parabolic antenna gives a much higher signal to noise ratio, but the transmitting and receiving antennas must be accurately aligned with each other.

Before fibre optics, for decades, these microwaves formed the heart of the long-distance telephone transmission system.

Microwaves do not pass through buildings well. In addition, even though the beam is well focused, there is still some divergence in space. Some waves may be refracted off low lying atmospheric layers and may take slightly longer to arrive than direct waves. Being out of phase they can cancel the signal. This effect is called multipath fading and is often a serious problem. It is weather and frequency dependent.

Bands up to 10 GHz are now in routine use, but at about 8 GHz a new problem sets in: absorption by water (rain). The only solution is to shut off links that are being rained on and route around them.

Microwave is also relatively inexpensive. Putting up two simple towers (maybe just big poles with four guy wires) and putting antennas on each one may be cheaper than burying 50 km of fibre through a congested urban area, and it may also be cheaper than leasing the telephone company fibre.

Microwaves have also another important use. We are speaking about cordless telephones, garage door openers, wireless hi-fi speakers, security gates etc. These devices use so called Industrial/Scientific/Medical bands forming an exception to the licensing rule: transmitters using these bands do not require government licensing. One band is allocated world-wide: 2.400-2.484 GHz. These bands are popular also for various forms of short-range wireless networking.

2.3.4. Infrared and Millimeter Waves

Unguided infrared and millimeter waves are widely used for short-range communication (remote control of televisions and stereos). They are relatively directional, cheap and easy to build, but they do not pass through the solid objects. For this reason, no government license is needed to operate an infrared system.

These properties have made infrared an interesting candidate for indoor wireless LANs (i.e. portable computers with infrared capability can be on local LAN without having to physically connect to it.

Infrared communication cannot be used outdoors because the sun shines as brightly in the infrared as in visible spectrum.

2.3.5. Lightwave Transmission

Unguided optical signaling has been in use for centuries.

A modern application is to connect the LANs in two building via lasers mounted on their rooftops. Optical signaling using lasers is unidirectional, so each building needs its own laser and its own photodetector. This scheme offers very high bandwidth and very low cost. It is also relatively easy to install and does not require license.

The laser’s strength, a very narrow beam, is also a weakness here. Aiming a laser beam 1 mm wide at a target 1 mm wide 500 m away could be a problem. Usually, lenses are put into the system to defocus the beam slightly.

A disadvantage is that laser beams cannot penetrate rain or thick fog. Some other phenomena in the atmosphere can also influence the communication using laser (Fig. 2-13.).

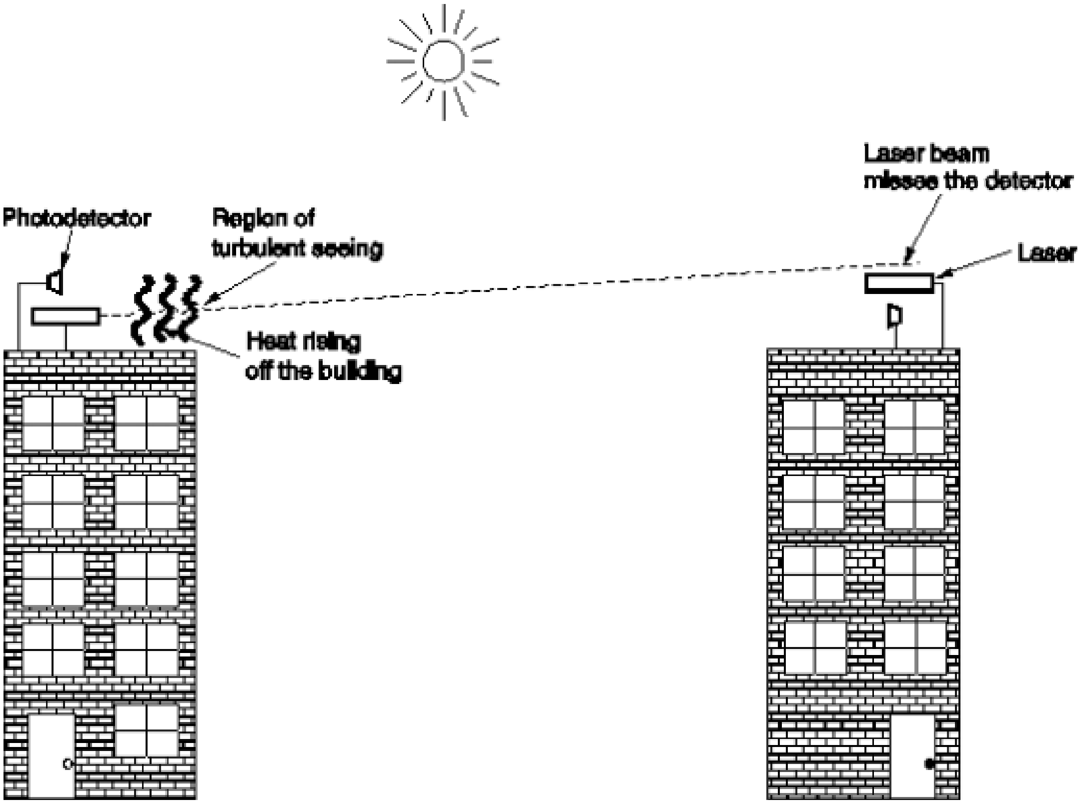


Fig. 2-13. Convection currents can interfere with laser communication systems. A bidirectional system, with two lasers, is pictured here.

2.4. The Telephone System

The telephone system is tightly intertwined with (wide area) computer networks, so it is worth to study it.

Public Switched Telephone Network (PSTN), that is used today also for computer networking, was designed many years ago with a completely different goal in mind: to transmit the human voice in a more or less recognizable form. Its suitability for use in computer - computer communication is often marginal at best, but the situation is rapidly changing with the introduction of fibre optics and digital technology.

An example illustrating the magnitude of the problem: Comparing a standard error rate of the cable connection of two computers running at memory speeds (107 to 108 bps), say 1 error per day, with the standard error rate when the computers communicate through a dial-up line at 104 bps, that is about 1 per 105 bits sent, we get difference about 11 orders.

2.4.1. Structure of the Telephone System

The telephone was patented by Graham Bell in 1876. Initially, the telephones were sold in pairs and it was up to customer to string a single wire between them. The electrons returned through the earth.

Bell formed also the Bell Telephone Company which opened its first switching office in New Haven, Connecticut, in 1878. To make a call, the customer would crank the phone to ring in the telephone company office where the operator manually connected the caller to the callee using a jumper cable. (Fig. 2-14(b)).

Later, the switching offices had to be connected to make long-distance calls possible. Therefore second-level switching offices became necessary (Fig. 2-14(c)) and successively the hierarchy grew to five levels. This scheme remained essentially intact for over 100 years.

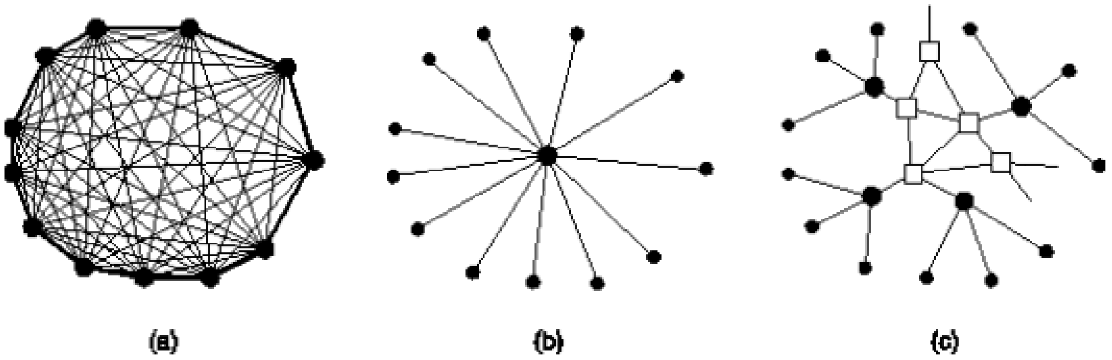


Fig. 2-14. (a) Fully interconnected network. (b) Centralized network. (c) Two level hierarchy.

At present, the telephone system can be, with some simplifications, described as follows: Each telephone has two copper wires coming out of it that go directly to the telephone company's nearest end office (in the US there are about 19000 end offices). The two wire connection of the telephone and end office is called local loop.

If a subscriber attached to a given end office calls another subscriber attached to the same end office, the switching mechanism within the office sets up a direct electrical connection between the two local loops that remains intact for the duration of the call.

If a called telephone is attached to another end office, the path will have to be established somewhere higher up in the hierarchy. There are toll offices, primary, sectional, and regional offices that form a network by which the end offices are connected. They communicate with each other via high bandwidth interoffice trunks formed today by coaxial cables, microwaves and especially fiber optics. The number of different kinds of switching centers and their topology varies from country to country depending on its telephone density (Fig. 2-15.)

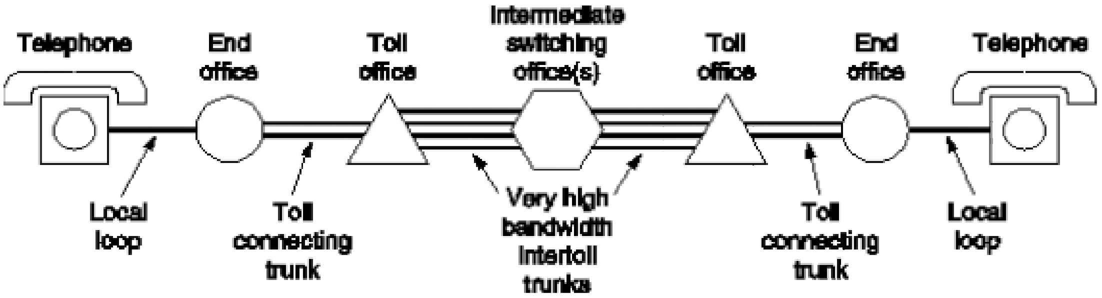


Fig. 2-15. Typical circuit route for a medium-distance call.

In the past, signaling throughout the telephone system was analog. Now, all the long-distance trunks within the telephone system are rapidly being converted to digital using optical fibers. It has the following reasons:

- Digital signal can pass through arbitrary number of regenerators with no information loss. In contrast, analog signals always suffer some information loss when amplified, and this loss is cumulative.

- Voice, data, music, and images can be interspersed to make more efficient use of the circuits and equipment.
- Much higher data rates are possible.
- Digital transmission is much cheaper than analog, since it is not necessary to accurately reproduce an analog waveform through potentially hundreds of amplifiers on a transcontinental call.
- Maintenance of digital system is easier. A transmitted bit is either received correctly or not.

In summary, the telephone system consists of three major components:

1. Local loops (twisted pairs, analog signaling).
2. Trunks (fiber optics or microwave, mostly digital).
3. Switching offices.

2.4.2. The Local Loop

The local loops are sill analog. Consequently, when a computer wishes to send digital data over a dial-up line, the data must first be converted to analog form by a modem for transmission over a local loop, then converted to digital form for transmission over the long-haul trunks, then back to analog over the local loop at the receiving end, and finally back to digital by another modem for storage in the destination computer (Fig. 2-17.).

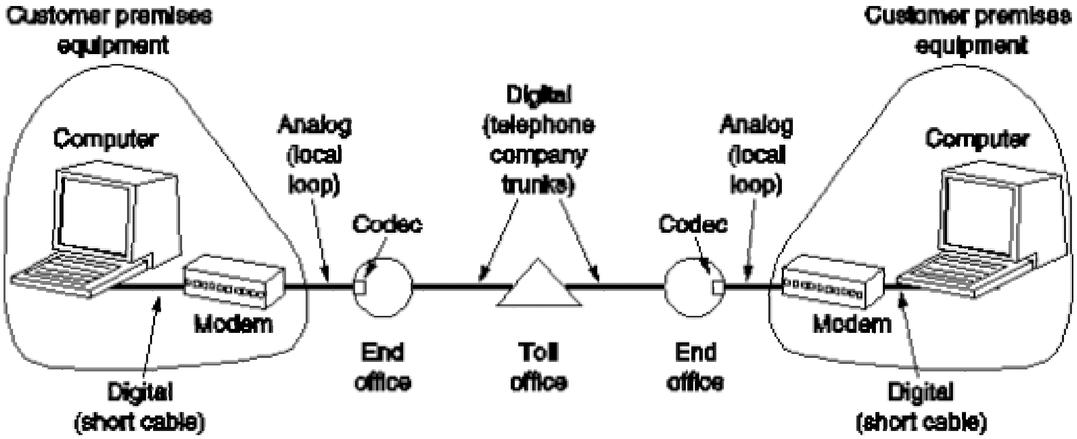


Fig. 2-17. The use of both analog and digital transmission for a computer to computer call. Conversion is done by the modems and codecs.

For leased lines it is possible to go digital from start to finish, but these lines are still expensive.

2.4.3. Transmission Impairments

Transmission lines suffer from three major problems:

1. *Attenuation*. It is the loss of energy as the signal propagates outwards. On guided media the signal falls off logarithmically with the distance. The loss is expressed in decibels per km. The amount of energy lost depends of frequency. Amplifiers can be put in to try to compensate for frequency-dependent attenuation. They help but can never restore the signal exactly back to its original shape.
2. *Delay distortion*. It is caused by the fact that different Fourier components travel at different speeds. For digital data, fast components from one bit may catch up and overtake slow components from the bit ahead, mixing the two bits and increasing the probability of incorrect reception.

3. *Noise*. It is unwanted energy from sources other than transmitter (thermal noise, cross talks, impulse noise).

2.4.4. Modems

Due to transition impairments dependent on frequency, it is undesirable to have a wide range of frequencies in the signal. Square waves of digital data have a wide spectrum and thus are subject to strong attenuation and delay distortion. So the baseband (DC) signaling is unsuitable except at slow speed and over short distances.

To get around the problem, especially on telephone lines, analog (AC) signaling is used. It is based on continuous tone in the 1000 to 2000 Hz range, called sine wave carrier, with amplitude, frequency or phase modulation to transmit information (Fig. 2-18.).

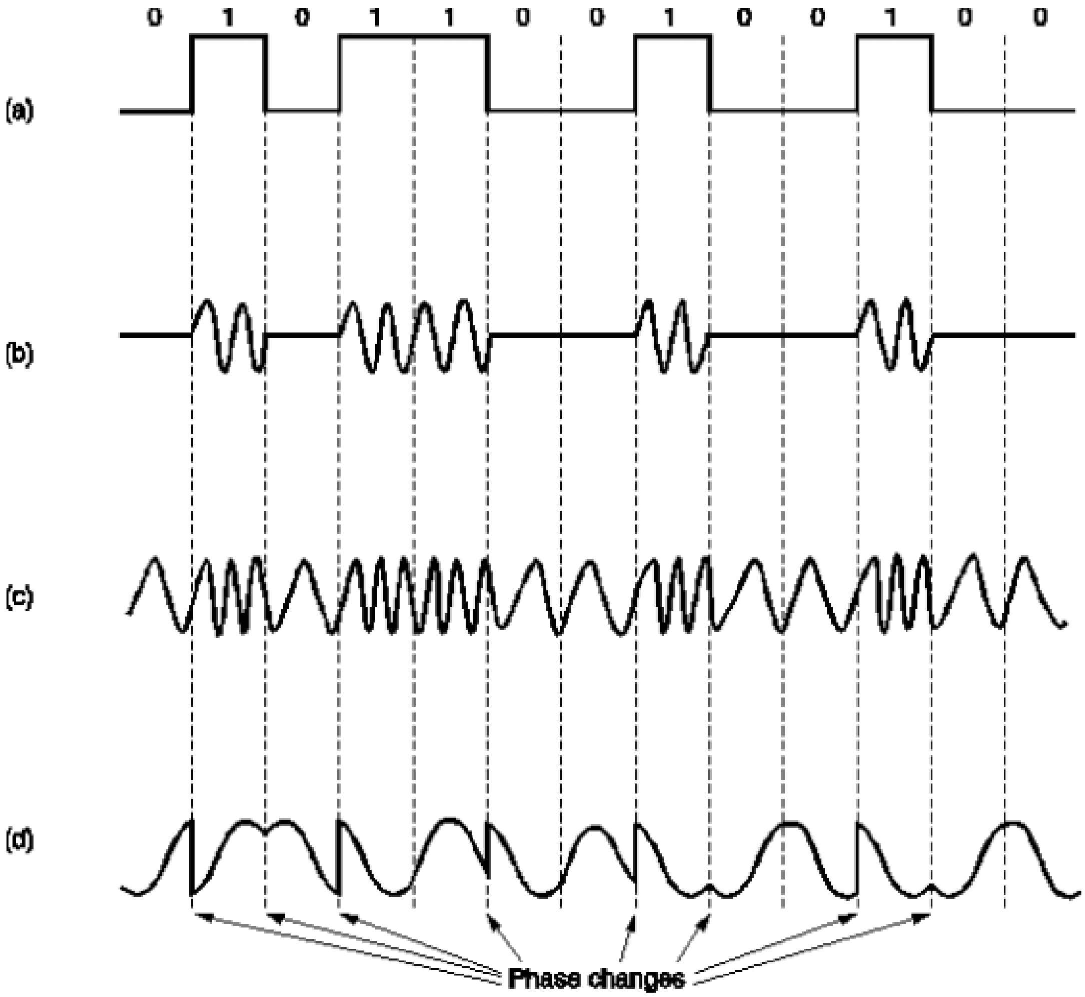


Fig. 2-18. (a) A binary signal. (b) Amplitude modulation. (c) Frequency modulation. (d) Phase modulation.

In amplitude modulation, two different voltage levels are used to represent 0 and 1, respectively.

In frequency modulation (or frequency shift keying), two or more different tones are used.

In phase modulation, the carrier wave is systematically shifted at uniformly spaced intervals. E.g., if 45, 135, 225, or 315 degrees shifts are used, each phase shift transmits 2 bits of information.

A device that accepts a serial stream of bits as input and produces a modulated carrier as output (and vice versa) is called *modem* (for modulator-demodulator).

To go to the higher speeds, it is not possible to just keep increasing the sampling rate. The Nyquist theorem says that even with the perfect 3000 Hz line there is no point in sampling faster then 6000 Hz. Thus all research on faster modems is focused on getting more bits per sample (i.e. per baud).

Most advanced modems use a combination of modulation techniques to transmit multiple bits per baud. 2 combinations of amplitude levels and phase shifts are displayed in Fig. 2-19. Such diagrams are called constellation patterns and each high-speed modem standard has its own one. The ITU V.32 9600 bps modem standard uses the constellation pattern of Fig. 2-19(b).

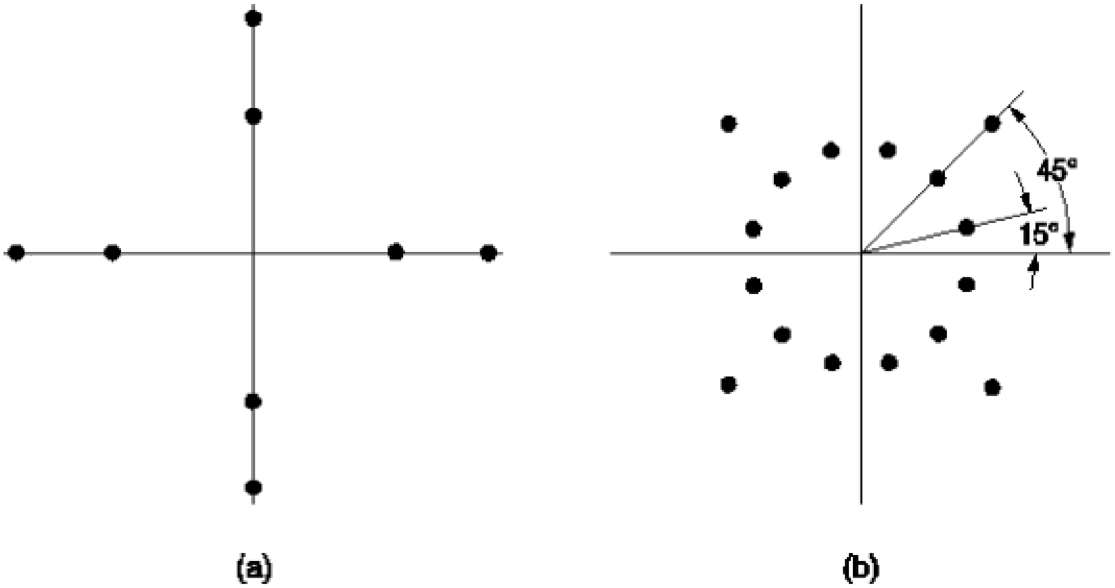


Fig. 2-19. (a) 3 bits/baud modulation. (b) 4 bits/baud modulation.

The next step above 9600 bps is 14400 bps (V.32 bis). This speed is achieved by transmitting 6 bits per sample at 2400 baud. Its constellation pattern has 64 points. After V.32 bis comes V.34 running at 28800 bps.

Many modems now have compression and error correction built into the modems. It improves the effective data rate. One popular compression scheme is MNP 5, which uses run-length encoding to squeeze out runs of identical bytes.

Full-duplex transmission. i.e. transmission in both directions at the same time is based on the use of different frequency bands for each direction. The alternative is *half-duplex transmission*, in which communication can go either way, but only one at a time.

2.4.5. RS-232-C and RS-449

The interface between the computer and the modem is an example of a physical layer protocol. It must specify in detail the mechanical, electrical, functional and procedural interface. Two well-known physical layer standards are RS-232-C and its successor, RS-449.

RS-232-C is the third revision of the original *RS-232* standard drawn up by the Electronic Industries Association (EIA). Its international version is given by CCITT recommendation V.24 and differs very slightly on some of rarely used circuits. In the standards, the terminal or computer is officially called DTE (Data Terminal Equipment) and the modem is officially called a DCE (Data Circuit-Terminating Equipment).

The mechanical specification is for a 25-pin connector 47.04+-.13 mm wide (screw center to screw center), with all the other dimensions equally well specified. The top row has pins numbered 1 to 13 (left to right); the bottom row has pins numbered 14 to 25 (also left to right).

The electrical specification for RS-232-C is that voltage more negative than -3 volts is a binary 1 and a voltage more positive than +4 volts is a binary 0. Data rates up to 20kbps are permitted, as are cables up to 15 meters.

The functional specification tells which circuits are connected to each of the 25 pins, and what they mean. Fig. 2-21 shows 9 pins that are nearly always implemented. The remaining ones are frequently omitted.

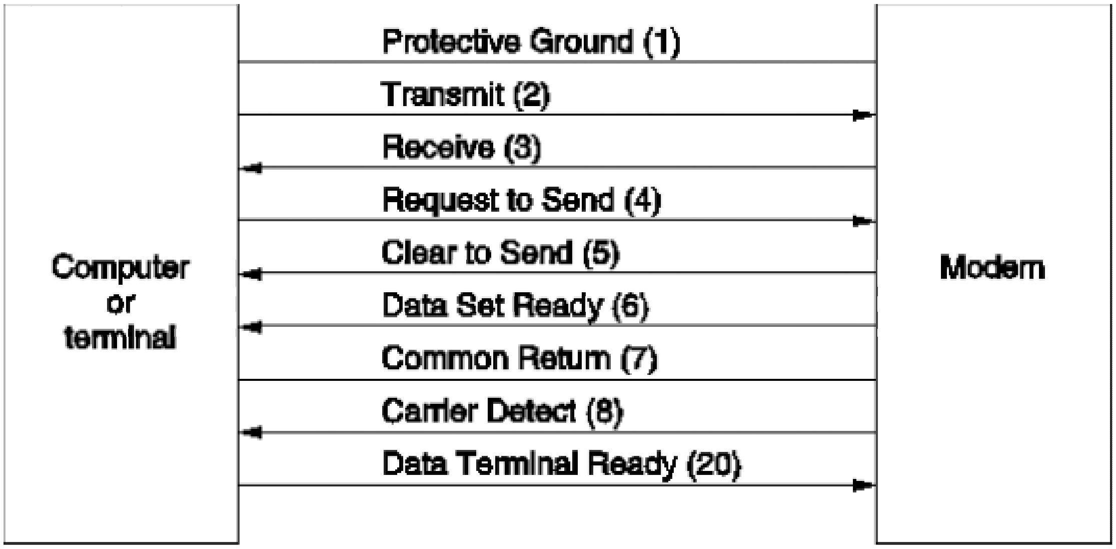


Fig. 2-21. Some of the pricipal RS-232-C circuits. The pin numbers are given in parentheses.

The specification of some signals:

- When the computer is powered up, it asserts (i.e. sets to logical 1) Data Terminal Ready (pin 20).
- When the modem is powered up, it asserts Data Set Ready (pin 6).
- When the modem detects a carrier on the telephone line, it asserts Carrier Detect (pin 8).
- Request to Send (pin 4) indicates that the computer want to send data.
- Clear to Send (pin 5) means that the modem is prepared to accept data.
- Data are transmitted on the Transmit circuit (pin 2) and received on the Receive circuit (pin 3).

Other circuits are provided for selecting data rate, testing the modem, detecting ringing signal, etc. They are usually not used in practice.

The procedural specification is the protocol, that is, the legal sequence of events. The protocol is based on action-reaction pairs. When the computer asserts Request to Send, for example, the modem replies with Clear to Send, if it is able to accept data.

It commonly occurs that two computers must be connected using RS-232-C. Since neither is a modem, there is an interface problem. It is solved by connecting the computers with the device called null modem, which connects the transmit line of one machine to the receive line of the other. It also crosses some of the other lines in a similar way.

The successor of RS-232-C, RS-449, removes some of the limitations of RS-232-C. Among them, it can be used at speeds up to 2Mbps and over 60 meter cables.

2.4.6. Fiber in the Local Loop

For advanced future services, such as video on demand, the 3-kHz channel currently used will not do. Two possibilities of what to do are discussed:

1. Running a fiber from end office into everyone's house called FTTH *(Fiber To The Home)*. This solution fits in well with the current system but it is too expensive.
2. Running an optical fiber from each end office into each neighborhood (the curb) that it serves (FTTC - *Fiber To The Curb*). The fiber is terminated in a junction box that all the local loops enter. Since the local loops are now much shorter (around 100 m), they can be run at higher speeds, around 1 Mbps (Fig. 2-23(a)). An alternative design uses existing cable TV infrastructure (Fig. 2-23(b)).

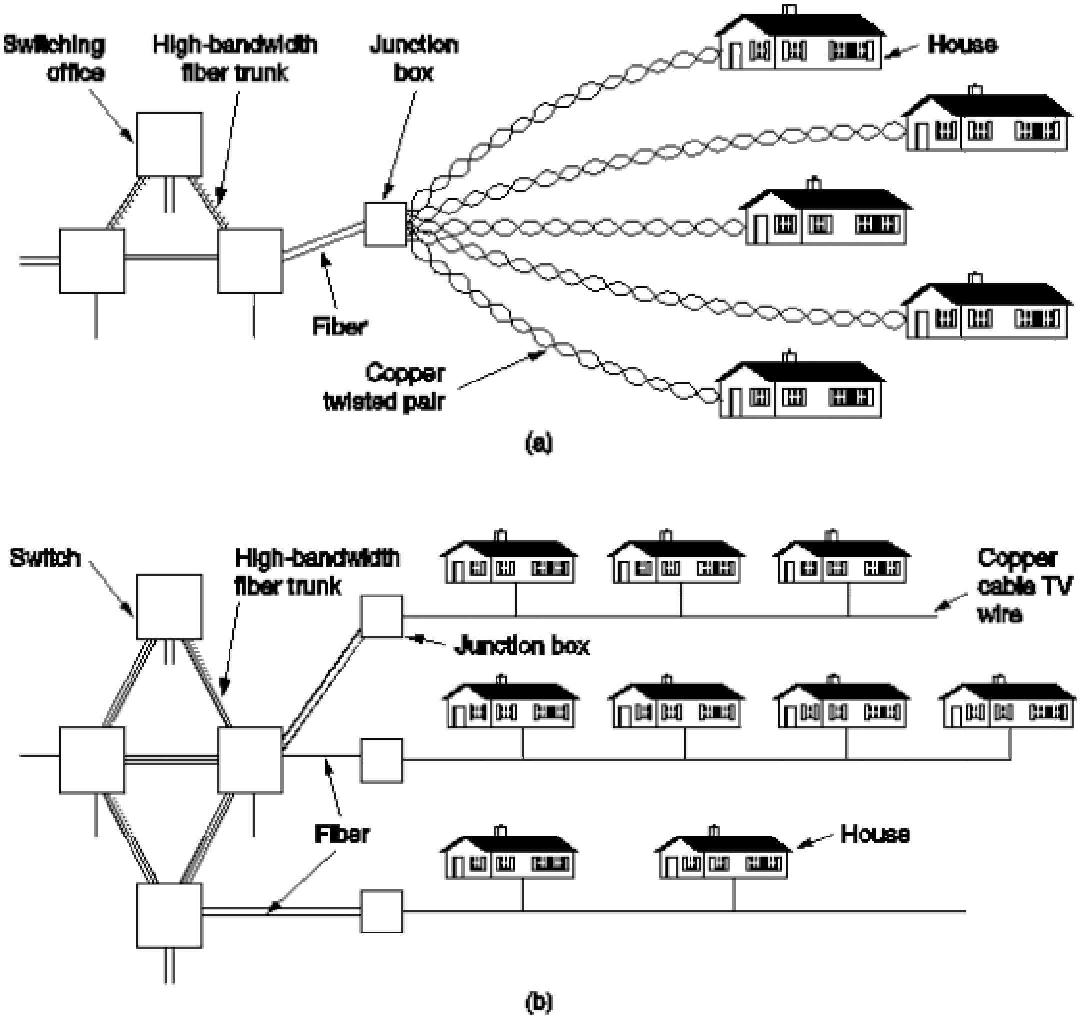


Fig. 2-23. Fiber to the curb. (a) Using the telephone network. (b) Using the cable TV network.

2.4.7. Trunks and multiplexing

Telephone companies have developed elaborate schemes for multiplexing many conversations over a single physical trounce. This multiplexing schemes can be divided into two basic categories:

1. *Frequency Division Multiplexing* (FDM) - the frequency spectrum is divided among the logical channels, single frequency bands are allocated to different users. As an example from another area of life, we can take radio broadcasting where different frequencies are allocated to different radio stations.
2. *Time Division Multiplexing* (TDM) - the users take turns (in a round robin), each one periodically getting the entire bandwidth for a little burst of time. Compare the burst of music alternated by the burst of advertising in radio broadcasting as an illustration.

2.4.8. Frequency Division Multiplexing

When 3000 Hz wide voice-grade telephone channels are multiplexed using FDM, 4000 Hz is allocated to each channel to keep them well separated. First, the voice channels are raised in frequency, each by a different amount, and then they are combined (Fig. 2-24).

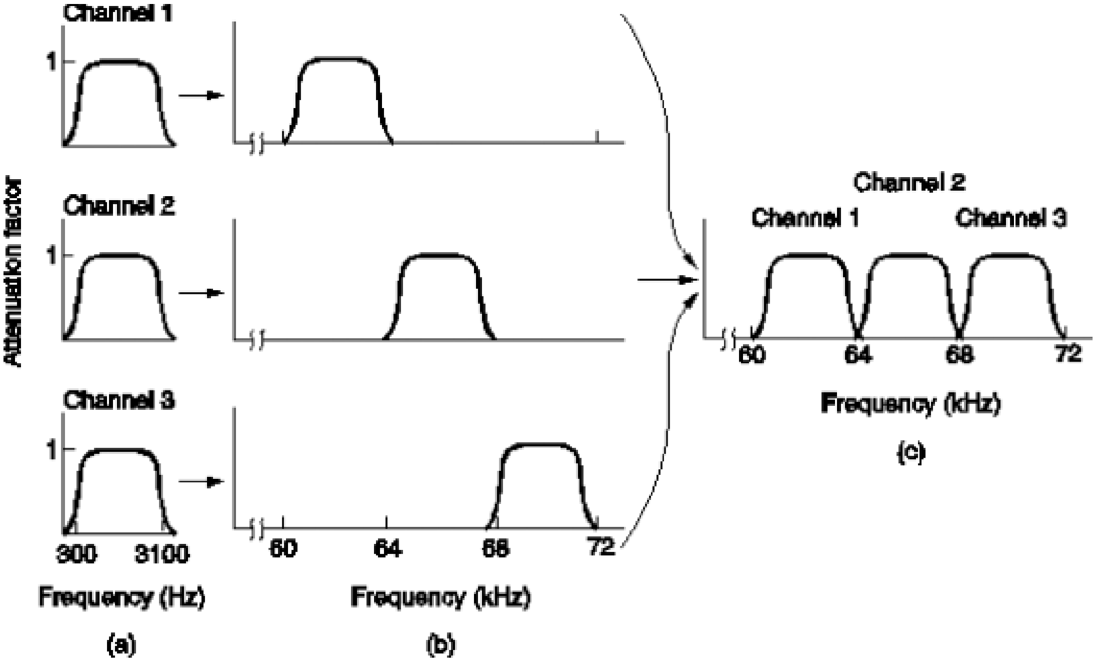


Fig. 2-24. Frequency division multiplexing. (a) The original bandwidths. (b) The bandwidths raised in frequency. (c) The multiplexed channel.

The FDM schemes used around the world are to some degree standardized. A widespread standard is 12 4000 Hz voice channels multiplexed into 60 to 108 kHz band. This unit is called a group. Five groups can be multiplexed to form a supergroup, five supergroups form a mastergroup.

For fiber optic channels, a variation of frequency division multiplexing called Wavelength Division Multiplexing (WDM) is used (Fig. 2-25).

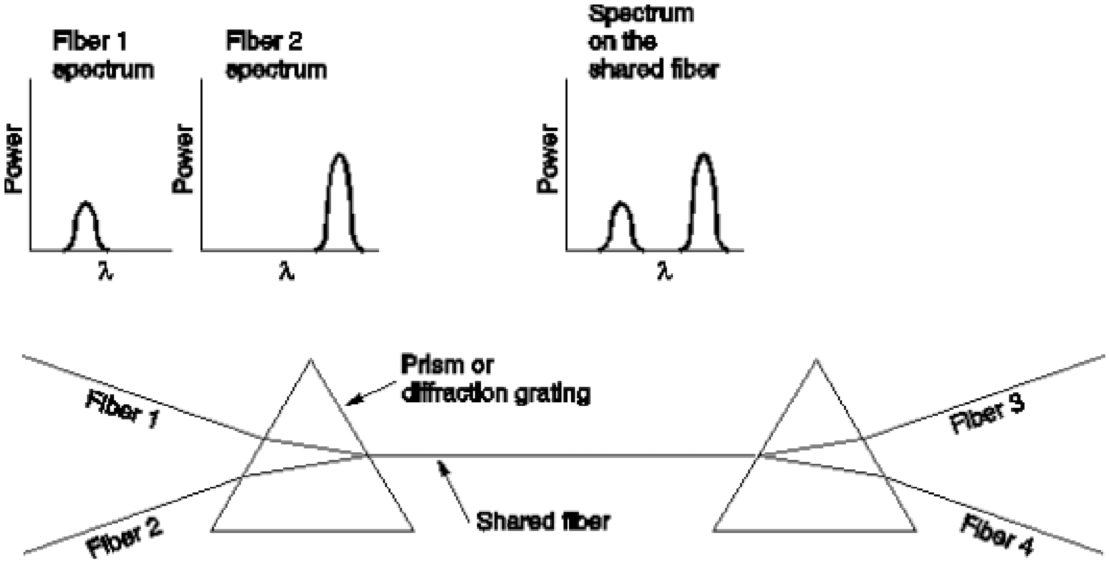


Fig. 2-25. Wavelength division multiplexing.

2.4.9. Time Division Multiplexing

FDM requires analog circuitry and cannot be performed by a computer. In contrast, TDM can be handled entirely by digital electronics, so it has become far more widespread in recent years. But it can only be used for digital data. Since the local loops produce analog signals, they must be converted in the end offices to be combined onto outgoing trunks.

The analog signals are digitized in the end office by a device called a *codec* (coder-decoder), producing a 7 or 8 bit number (see Fig. 2-17). The codec makes 8000 samples per second (125 (sec/sample) because the Nyquist theorem says that this is sufficient to capture all the information from the 4 kHz telephone channel bandwidth. This technique is called Pulse Code Modulation - PCM. PCM forms the heart of the modern telephone system.

There are a variety of incompatible schemes in use for PCM in different countries around the world. One method that is in widespread use in North America and Japan is the T1 carrier (Fig. 2-26). It consists of 24 voice channels multiplexed together. 24 analog signals are sampled on a round-robin basis during each 125 (sec interval each channel getting 8 bits (possibly 7 bits of data and one for control) into the output stream. The resulting frame contains 24 x 8 = 192 bits plus one extra bit for framing, yielding 193 bits every 125 (sec. This gives a gross data rate 1.544 Mbps. The 193rd bit used for synchronization takes on the pattern 010101010101 .. . When the T1 system is being used entirely for data, only 23 of the channels are used for data. The 24th one is used for a special synchronization pattern, to allow faster recovery in the event of error.

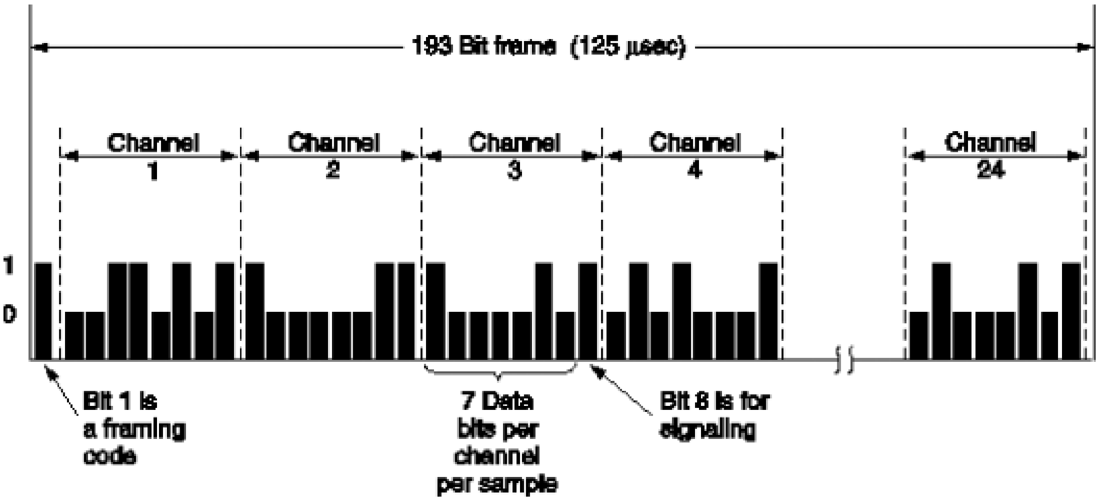


Fig. 2-26. The T1 carrier (1.544 Mbps).

There is also a CCITT recommendation for a PCM carrier at 2.048 Mbps called E1. This carrier has 32 8 bits data samples packed into the basic 125 (sec frame. This is in widespread use outside North America and Japan.

Once the voice signal has been digitalized, different compaction methods have been developed to reduce the member of bits needed per channel. All are based upon the principle that the signal changes relatively slowly compared to the sampling frequency, so that much of the information in the 7 or 8 bit digital level is redundant.

TDM allows multiple T1 carriers to be multiplexed into higher-order carriers (Fig. 2-28).

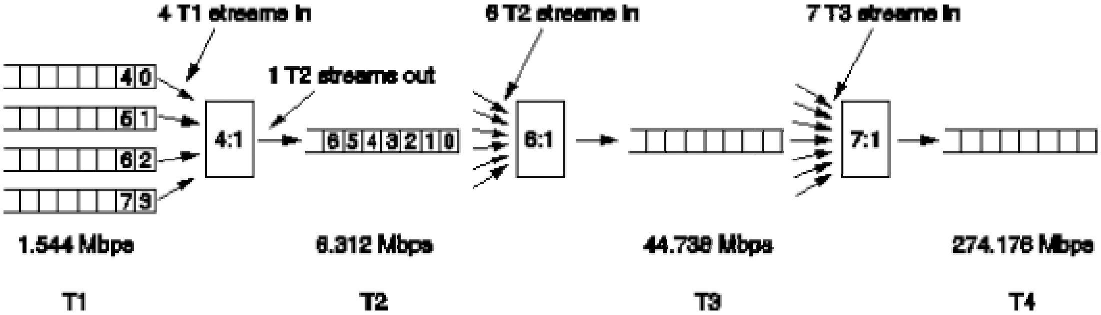


Fig. 2-28. Multiplexing T1 streams onto higher carriers.

2.4.10. SONET/SDH

After AT&T was broken up in 1984, the need for standardization for long-distance carriers became obvious. In 1985, Bellcore began working on a standard called SONET (Synchronous Optical NETwork). Later, CCITT joined the effort which resulted in SONET standard and a set of parallel CCITT recommendations (G.707, G.708, and G.709) in 1989. The CCITT recommendations are called SDH (Synchronous Digital Hierarchy) that differs from SONET only in minor ways. In fact all the long-distance telephone traffic in the US, and much of it elsewhere now using trunks running SONET in the physical layer. As SONET chips become cheaper, SONET interface boards for computers may become more widespread enabling to plug computers directly into the heart of the telephone network over special leased lines.

The SONET design had four major goals:

1. to make it possible for different carriers to interwork,

2. to unify U.S., European, and Japanese digital systems all working on the base of 64 kbps PCM channel but combining them in different ways,
3. to provide a way to multiplex multiple digital channel together. At the time SONET was devised, T4 was the highest channel as for speed. It was necessary to extend the scale higher.
4. to provide support for operation, administration and maintenance (OAM).

SONET is a synchronous system. It is controlled by a master clock. Bits on a SONET line are sent out at precise intervals, controlled by master clock.

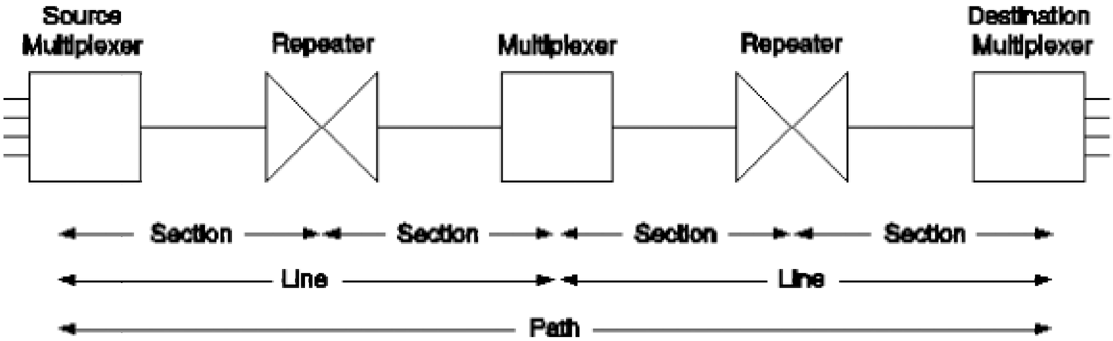


Fig. 2-29. A SONET path.

A SONET system consists of switches, multiplexers, and repeaters, all connected by fiber (Fig. 2-29). SONET terminology:

- section = a fiber going directly from any device to any other device, with nothing in between,
- line = a run between two multiplexers, possibly with one or more repeaters in the middle,
- path = a connection between the source and destination, possibly with one or more multiplexers and repeaters.

The basic SONET frame is a block of 810 bytes put out every 125 (sec. The 8000 frame/sec exactly matches the sampling rate of PCM channels used in all digital telephony systems.

The 810 byte SONET frames are best described as a rectangle of bytes, 90 columns wide by 9 rows high. The gross data rate is 8 x 810 x 8000 = 51.84 Mbps. This is a basic SONET channel called STS-1 (Synchronous Transport Signal -1). All SONET trunks are a multiple of STS - 1.

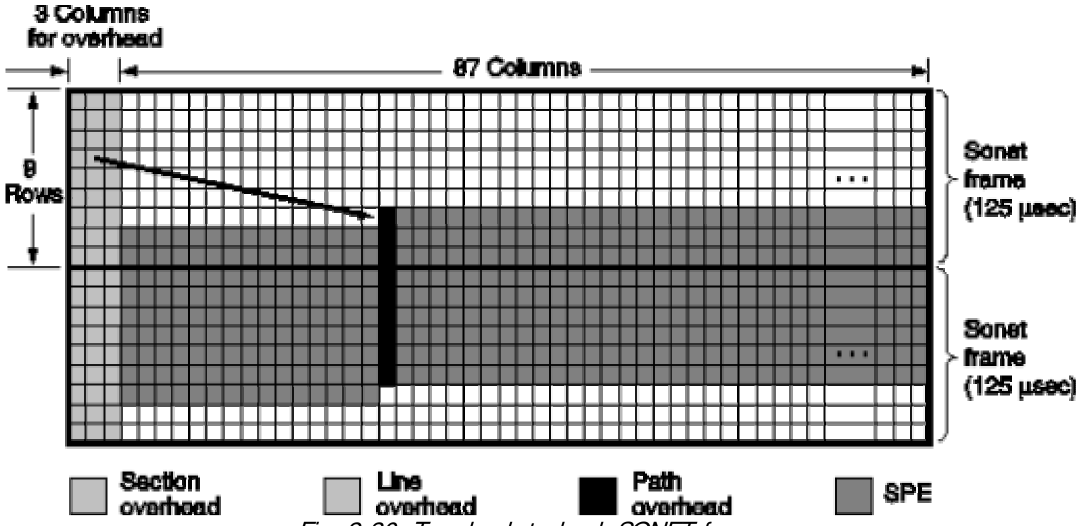


Fig. 2-30. Two back-to-back SONET frames.

The first three columns of each frame are reserved for system management information (Fig. 2-30). The first three rows contain the section overhead, the next six contain the line overhead.

The remaining 87 columns hold 50.112 Mbps of user data. However, the user data, called the Synchronous Payload Envelope - SPE - do not always begin in row 1, column 4. They can begin anywhere within the frame. A pointer to the first byte is contained in the first row of the line overhead. The first column of the SPE is the path overhead (i.e., header for the end-to-end path sublayer protocol).

The multiplexing of data streams, called tributaries, is illustrated in Fig. 2-31. The final output stream is STS-12 having 12 times the capacity of the STS - 1 stream. At this point the signal is scrambled, to prevent long runs of 0s or 1s from interfering with the clocking, and converted from an electrical to an optical signal. Multiplexing is done byte for byte.

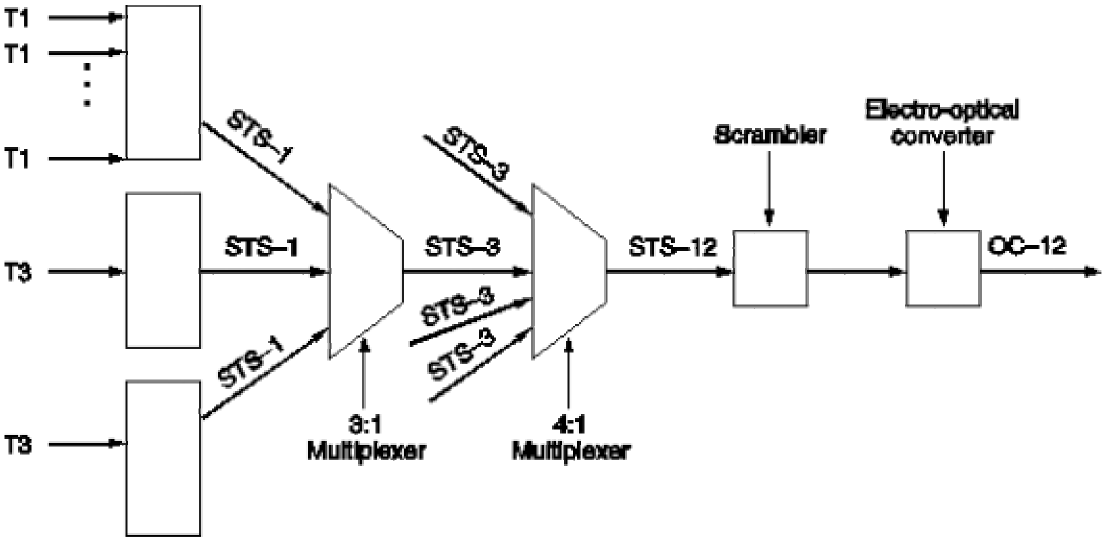
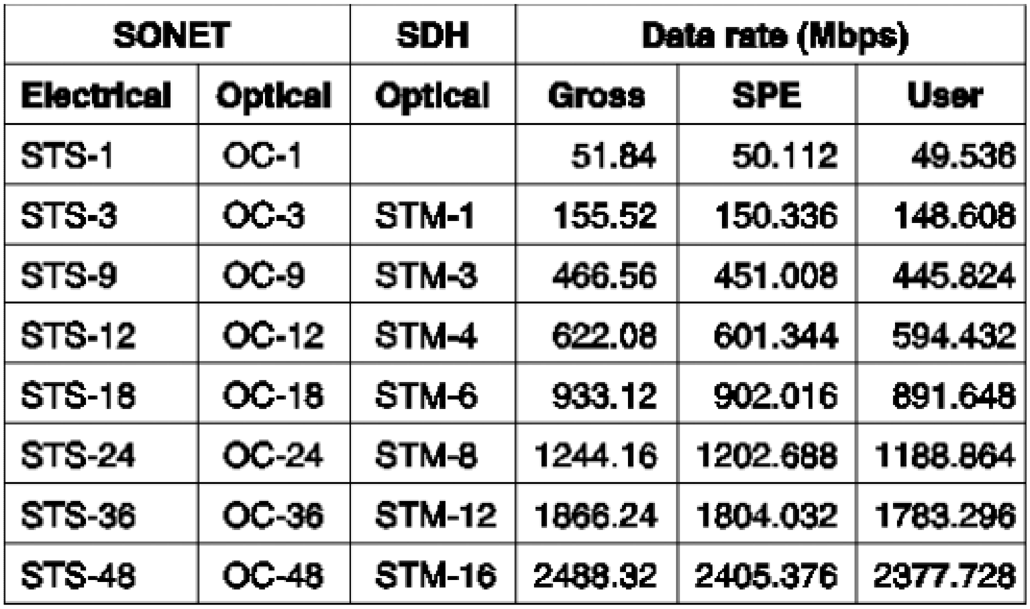


Fig. 2-31. Multiplexing in SONET.

The SONET multiplexing hierarchy is shown in Fig. 2-32. The optical carrier corresponding to STS-n is called OC-n. The SDH names are different, and they start at OC-3 because CCITT-based systems do not have a rate near 51.84 Mbps. The gross data rate includes all the overhead, the SPE data rate excludes the line and section overhead. The user data rate excludes all overhead.



SONET		SDH	Data rate (Mbps)		
Electrical	Optical	Optical	Gross	SPE	User
STS-1	OC-1		51.84	50.112	49.536
STS-3	OC-3	STM-1	155.52	150.336	148.608
STS-9	OC-9	STM-3	466.56	451.008	445.824
STS-12	OC-12	STM-4	622.08	601.344	594.432
STS-18	OC-18	STM-6	933.12	902.016	891.648
STS-24	OC-24	STM-8	1244.16	1202.688	1188.864
STS-36	OC-36	STM-12	1866.24	1804.032	1783.296
STS-48	OC-48	STM-16	2488.32	2405.376	2377.728

Fig. 2-32. SONET and SDH multiplex rates.

When a carrier, such as OC-3, is not multiplexed, but carries the data only from a single source, the letter c (concatenated) is appended to the designation. So OC-3c indicates a data stream from a single source at 155.52 Mbps. The amount of user data in an OC-3c stream is slightly higher than in OC-3 stream (149.760 Mbps versus 148.608 Mbps) because the path overhead column is included inside the SPE only once, instead of three times in case of three independent OC-1 streams.

By now it should be clear why ATM runs at 155 Mbps: the intention is to carry ATM cells over SONET OC-3c trunks. It should also be clear that the 155 Mbps is the gross rate, including the SONET overhead.

The SONET physical layer is divided up into four sublayers (Fig. 2-33):

- The photonic sublayer is concerned with specifying the physical properties of the light and fiber to be used.
- The section sublayer handles a single point-to-point fiber run, generating a standard frame at one end and processing it at the other.
- The line sublayer is concerned with multiplexing multiple tributaries onto a single line and demultiplexing them at the other end.
- The path sublayer and protocol deal with end-to-end issues.

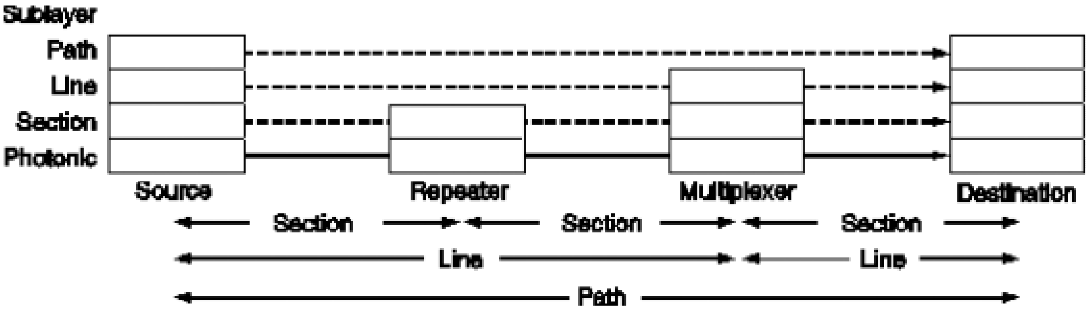


Fig. 2-33. The SONET architecture.

2.4.11. Switching

Two different switching techniques are used inside the telephone system:

- circuit switching
- packet switching

2.4.12. Circuit Switching

When a user place a telephone call, the switching equipment within the telephone system seeks out a physical "copper" (including fiber and radio) path from the caller telephone to the callee telephone. This technique is called circuit switching (Fig. 2-34(a)).

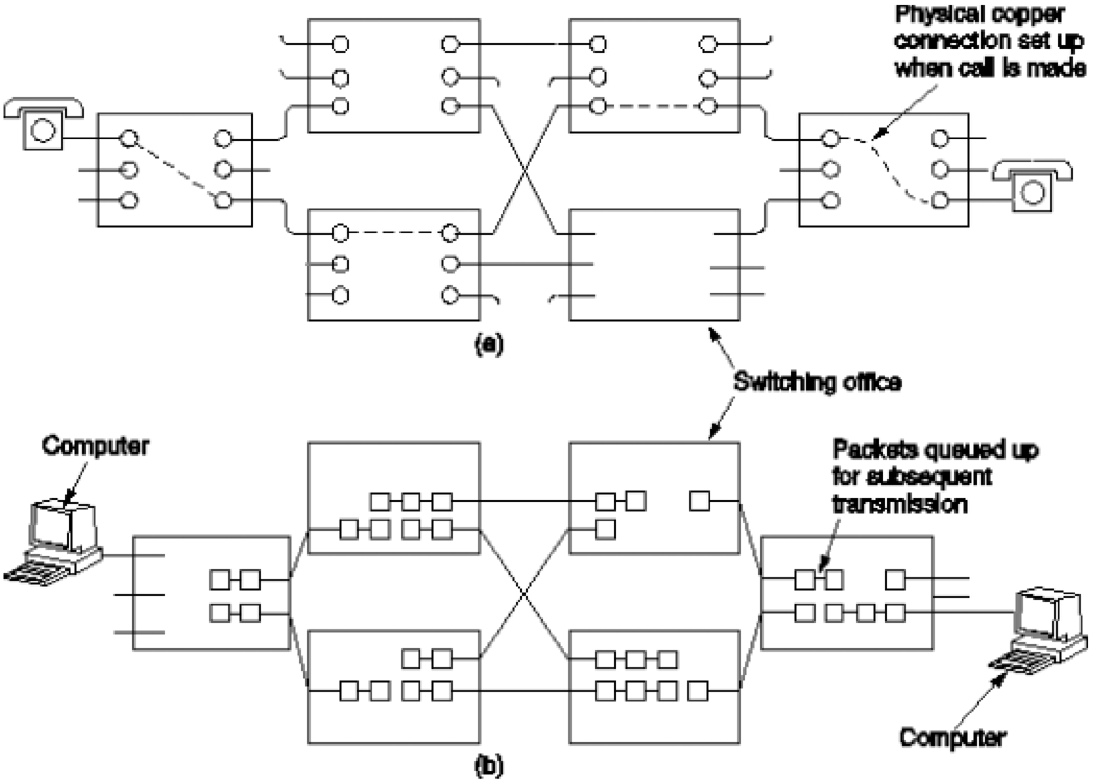

Physical copper connection set up when call is made
(a)
Switching office
Computer
Packets queued up for subsequent transmission
Computer
(b)

Fig. 2-34. (a) Circuit switching. (b) Packet switching.

An important property of circuit switching is the need to set up an end-to-end path before any data can be sent. It takes some set-up time during which there is no data transmission in progress. Long set-up times are for many computer applications undesirable.

As a consequence of the path between the calling parties, once the set-up has been completed, the only delay for data is the propagation time for the signal and there is no danger of congestion.

An alternative switching strategy is message switching (Fig 2-35(b)). In this case, no physical copper path is established in advance. Instead, the store-and-forward technique for the entire messages is applied. It was first used for telegrams.

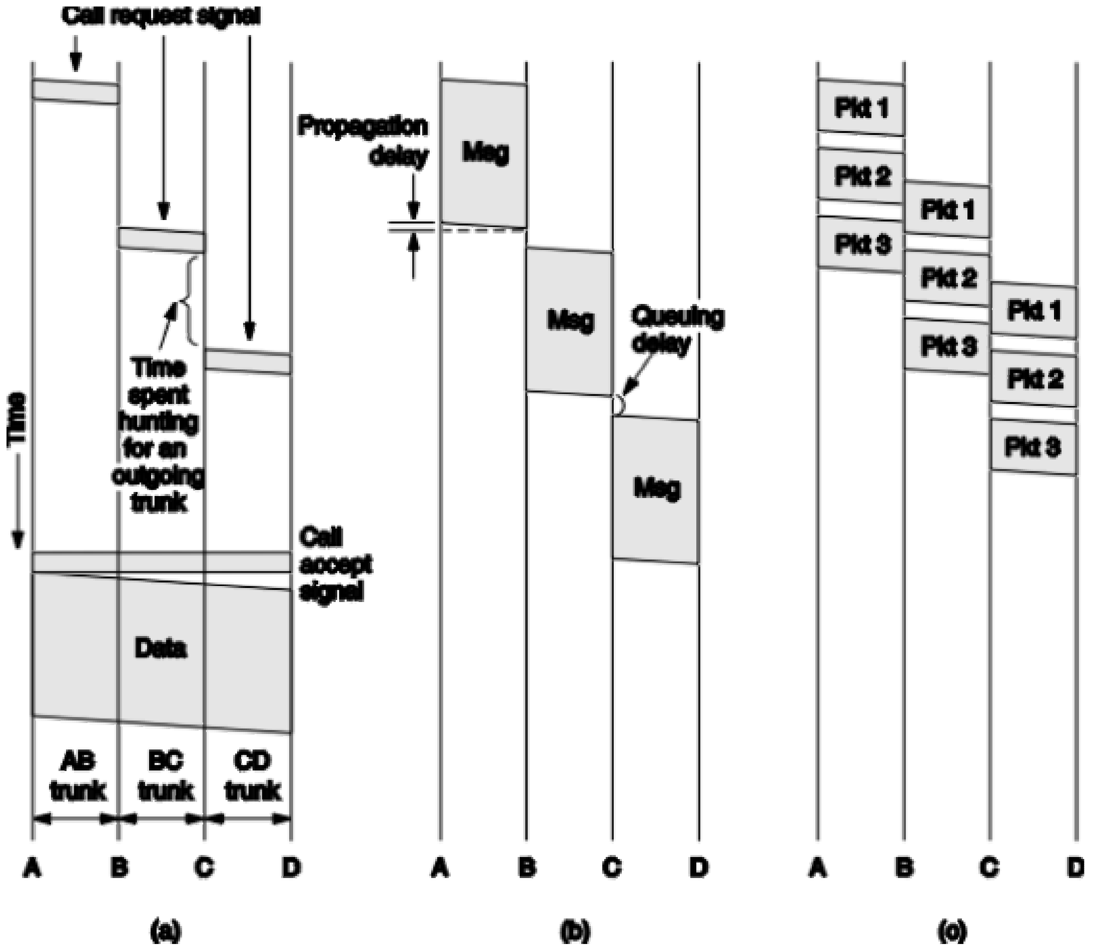
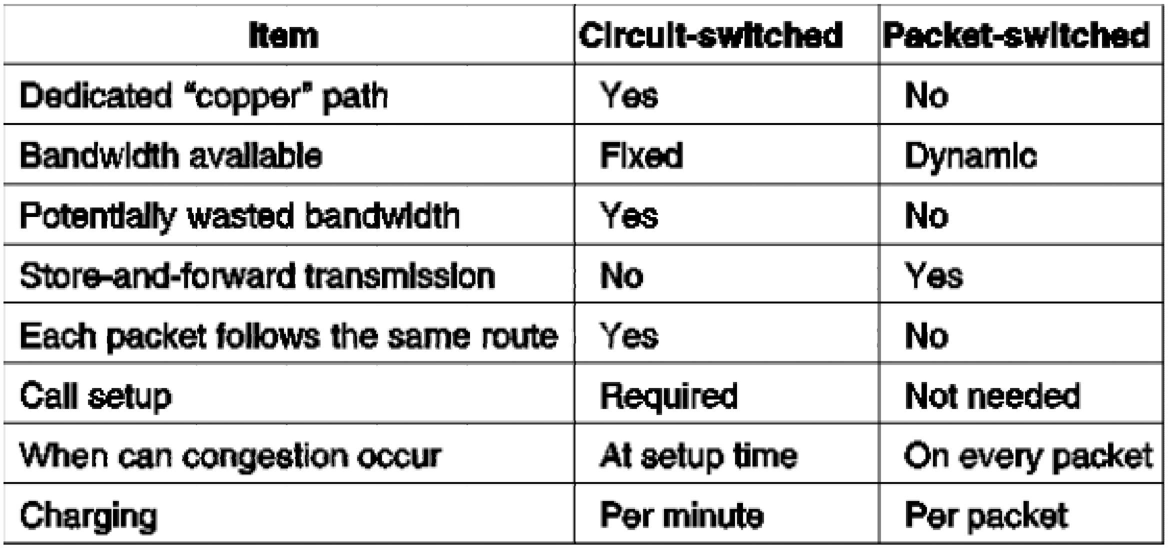


Fig. 2-35. Timing of events in (a) circuit switching, (b) message switching, (c) packet switching.

With message switching, there is no limit on block size, which means that routers must have disks to buffer long blocks. It also means that a single block may tie up a router-router line for minutes, rendering message switching useless for interactive traffic.

To get around these problems, packet switching was invented. Packet-switching networks place a tight upper limit on block size. So no user can monopolize any transmission line very long and therefore these networks are well suited to handle interactive traffic. A further advantage of packet switching over message switching is (Fig. 2-35(c)) that the first packet of a multipacket message can be forwarded before the second has fully arrived, reducing delay and improving throughput. For these reasons, computer networks are usually packet switched, occasionally circuit switched, but never message switched.

The differences between circuit switching and packet switching are summarized in Fig. 2-36.



Item	Circuit-switched	Packet-switched
Dedicated "copper" path	Yes	No
Bandwidth available	Fixed	Dynamic
Potentially wasted bandwidth	Yes	No
Store-and-forward transmission	No	Yes
Each packet follows the same route	Yes	No
Call setup	Required	Not needed
When can congestion occur	At setup time	On every packet
Charging	Per minute	Per packet

Fig. 2-36. A comparison of circuit-switched and packet-switched networks.

2.4.13. The Switch Hierarchy

As an example of circuit-switched telephone system we will briefly describe the AT&T system. The system of other companies or countries have the same general principles.

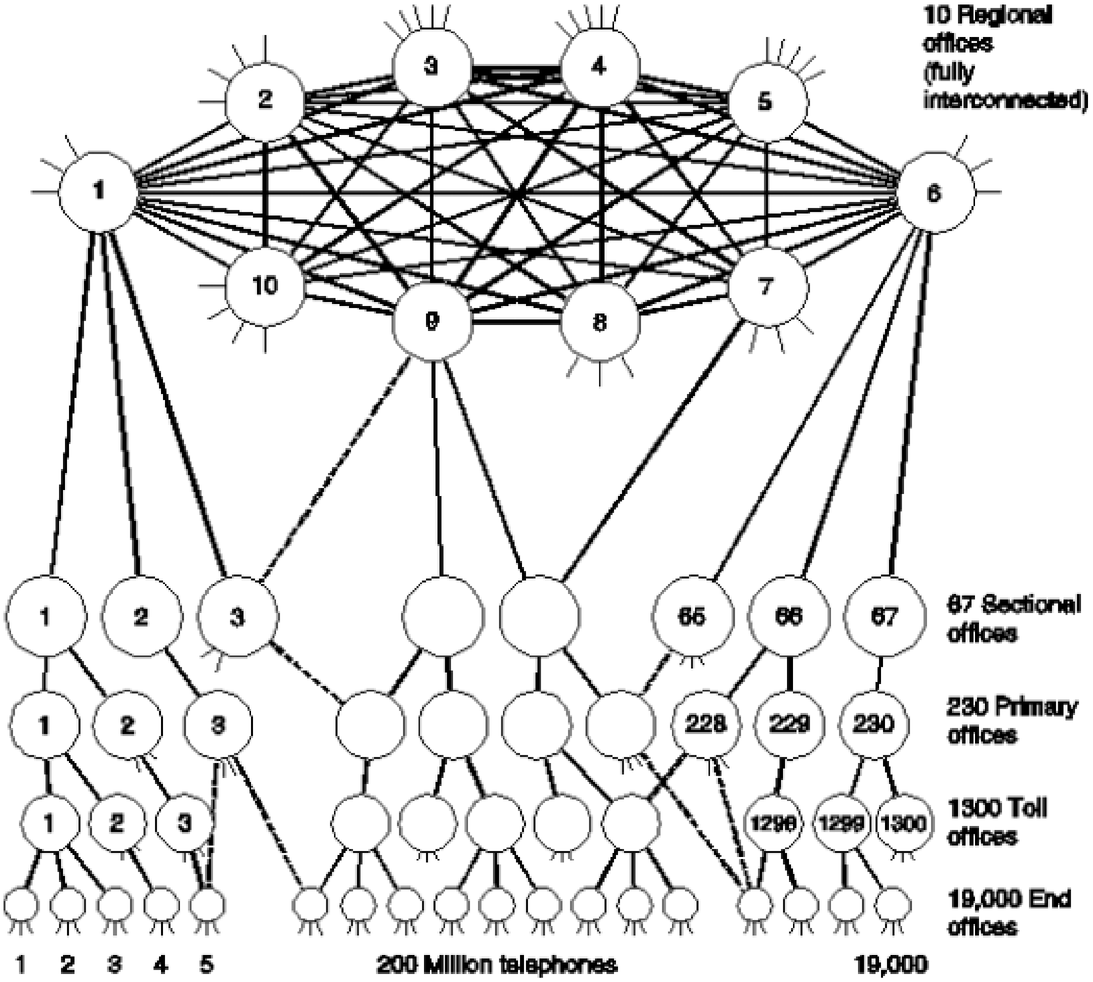


Fig. 2-37. The AT&T telephone hierarchy. The dashed lines are direct trunks.

The basic rules of operation of the system are (Fig. 2-37):

- the system has five classes of switching offices,
- calls are generally connected at the lowest possible level,
- some direct trunks for busy routes are installed - some calls can be routed along several paths.

2.4.14. Crossbar Switches

The *crossbar switch* (Fig.2-38) is the simplest kind of switch. In case of n input lines and n output lines it has n^2 crosspoints, where input and output lines can be connected by semiconductor switches.

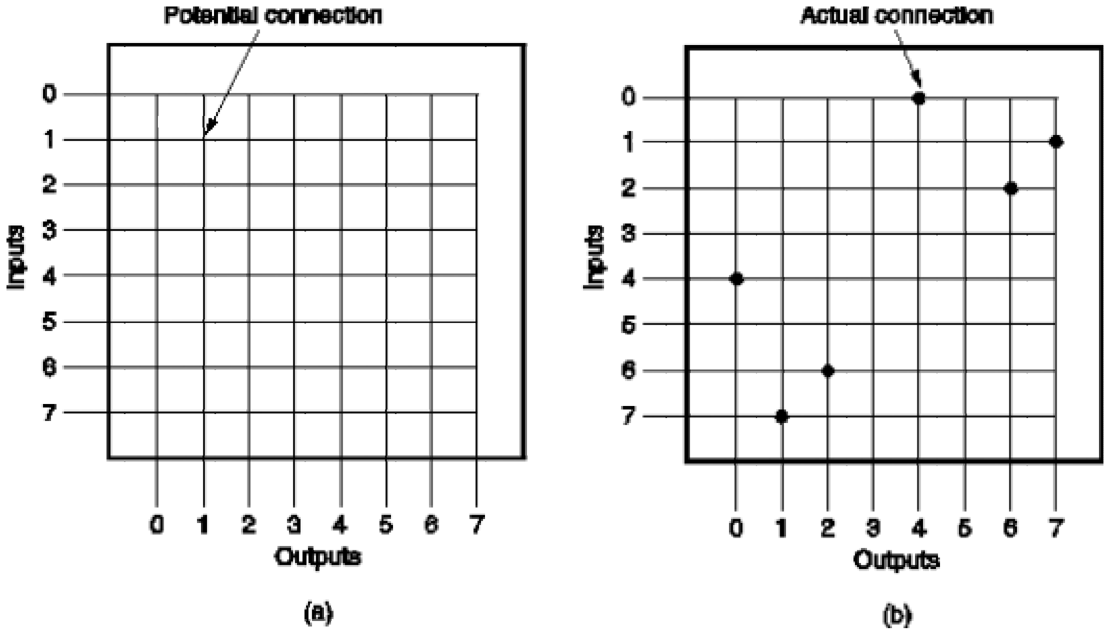


Fig. 2-38. A crossbar switch with no connections. (b) A crossbar switch with three connections set up: 0 with 4, 1 with 7, and 2 with 6.

The problem with a crossbar switch is that the number of crossbars grows as the square of the number of lines into the switch. If we assume that all lines are full duplex and that there are no self connections, only the crosspoint above the diagonal are needed. Still, n(n - 1)/2 crosspoints are needed. For n=1000, we need 499500 crosspoints. It possible to build a VLSI chip with this number of transistor switches, but not with 1000 pins on the chip. Thus a single crossbar switch is only useful for relatively small end offices.

2.4.15. Space Division Switches

By splitting the crossbar switch into smaller ones and interconnecting them, it is possible to build multistage switches with fewer crosspoints. These are called space division switches. Two configurations are illustrated in Fig. 2-39.

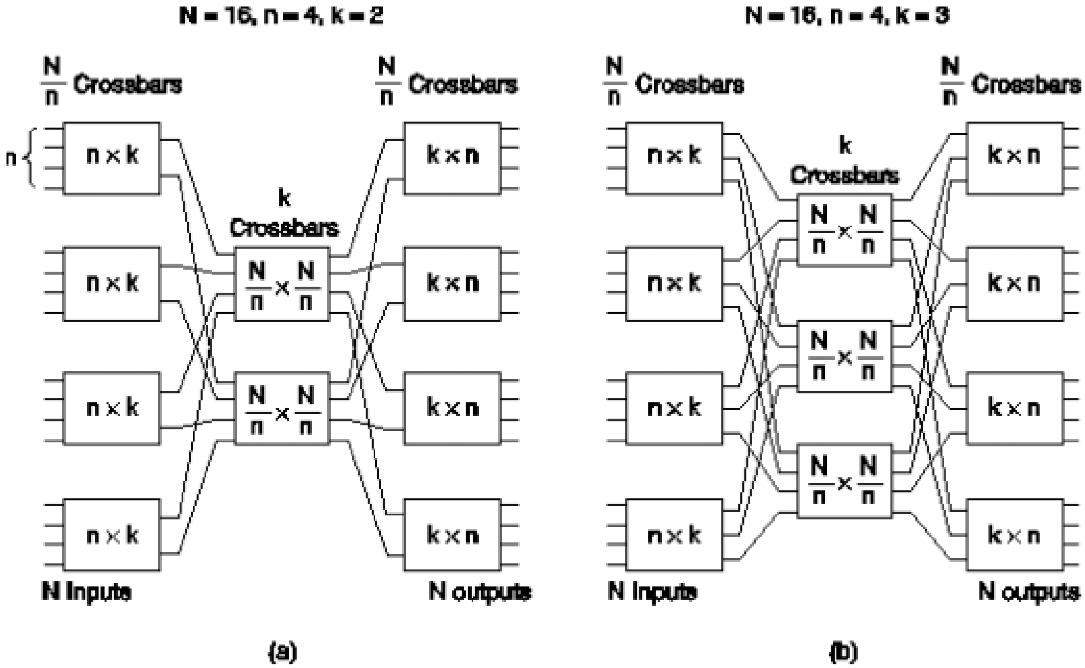


Fig. 2-39. Two space division switches with different parameters.

The number of crosspoints needed for a three-stage switch is 2kN + k(N/n)^2. For N = 1000, n = 50, and k = 10, we need only 24000 crosspoints instead of the 499500 required by a single-stage crossbar.

However, this type of switch can make much less connections at the same time comparing with the single-stage crossbar (8 in case (a), 12 in case (b)).

2.4.16. Time Division Switches

With time division switch (Fig. 2-40), the n input lines are scanned in sequence to build up an input frame with n slots. Each slot has k bits. For T1 switches, the slots are 8 bits, with 8000 frames processed per second.

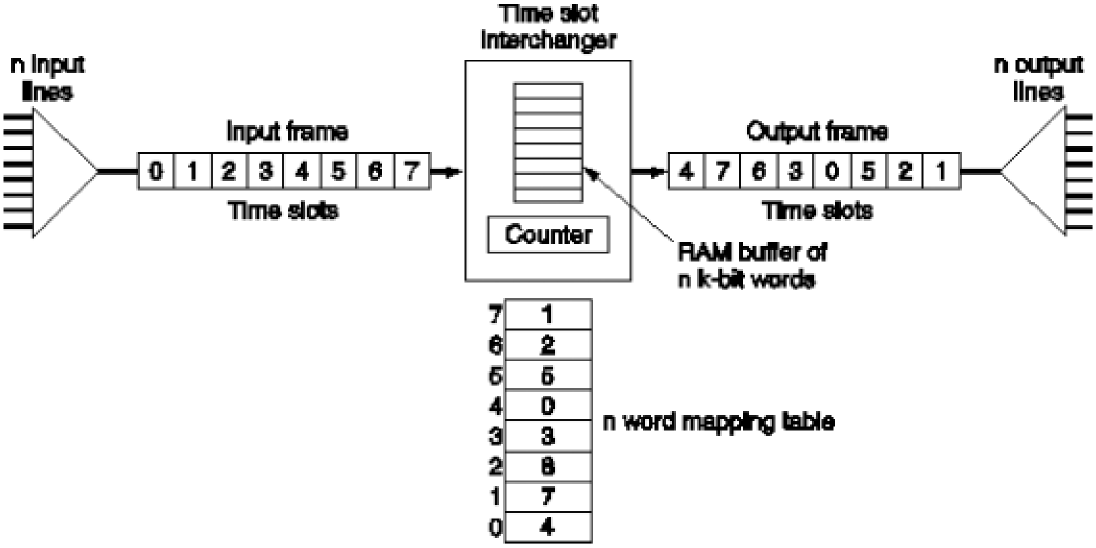


Fig. 2-40. A time division switch.

The heart of the time division switch is the time slot interchanger, which accepts input frames and produces output frames, in which the time slots have been reordered according to mapping table in the memory of the switch. Finally, the output frame is demultiplexed with output slot 0 going to line 0, and so on. In essence, the switch moves data from input lines to output lines according to the mapping table even though there are no physical connections between these lines.

The problem that limits the number of input lines to a time division switch is the time necessary to transform an input frame into the corresponding output frame. It is necessary to store n slots in the buffer RAM and then to read them out again within one frame period of 125 (sec. With memory access time T, we need a time interval 2nT, so with T = 100 nsec we can support at most n = 125/2T = 625 lines.

2.5. Narrowband ISDN

Anticipating user demand for end-to-end digital services the world's telephone companies agreed in 1984 under the auspices of CCITT to build a new, fully digital, circuit-switched telephone system by the early part of the 21st century. This system was called ISDN (Integrated Services Digital Network) and its primary goal was to integrate the voice and nonvoice services. It is already available in many locations and its use is growing slowly.

2.5.1. ISDN Services

The key ISDN service will continue to be voice but with many enhanced features.

Some of them are:

- buttons for instant call setup to arbitrary telephones anywhere in the world,
- displaying the caller's telephone number, name and address while ringing,
- connecting the telephone to a computer enabling the caller's database record to be displayed on the screen as the call comes in,
- call forwarding,
- conference calls worldwide,
- on line medical, burglar, and smoke alarms giving the address to speed up response.

2.5.2. ISDN System Architecture

The key idea behind ISDN is that of the digital bit pipe between the customer and the carrier through which bits flow in both directions. Whether the bits originate from a digital telephone, a digital terminal, a digital facsimile machine, or some other device is irrelevant.

The digital bit pipe can support multiple independent channels by time division multiplexing of the bit stream. Two principal standards for the bit pipe have been developed:

- a low bandwidth standard for home use, and
- a higher bandwidth standard for business use that supports multiple channels identical to the home use channels.

Normal configuration for a home consists of a network terminating device NT1 (Fig. 2-41(a)) placed on the customer's premises and connected to the ISDN exchange in the carrier's office using the twisted pair previously used to connect the telephone. The NT1 box has a connector into which a bus cable can be inserted. Up to 8 ISDN telephones, terminals, alarms, and other devices can be connected to the cable. From the customer's point of view, the network boundary is the connector on NT1.

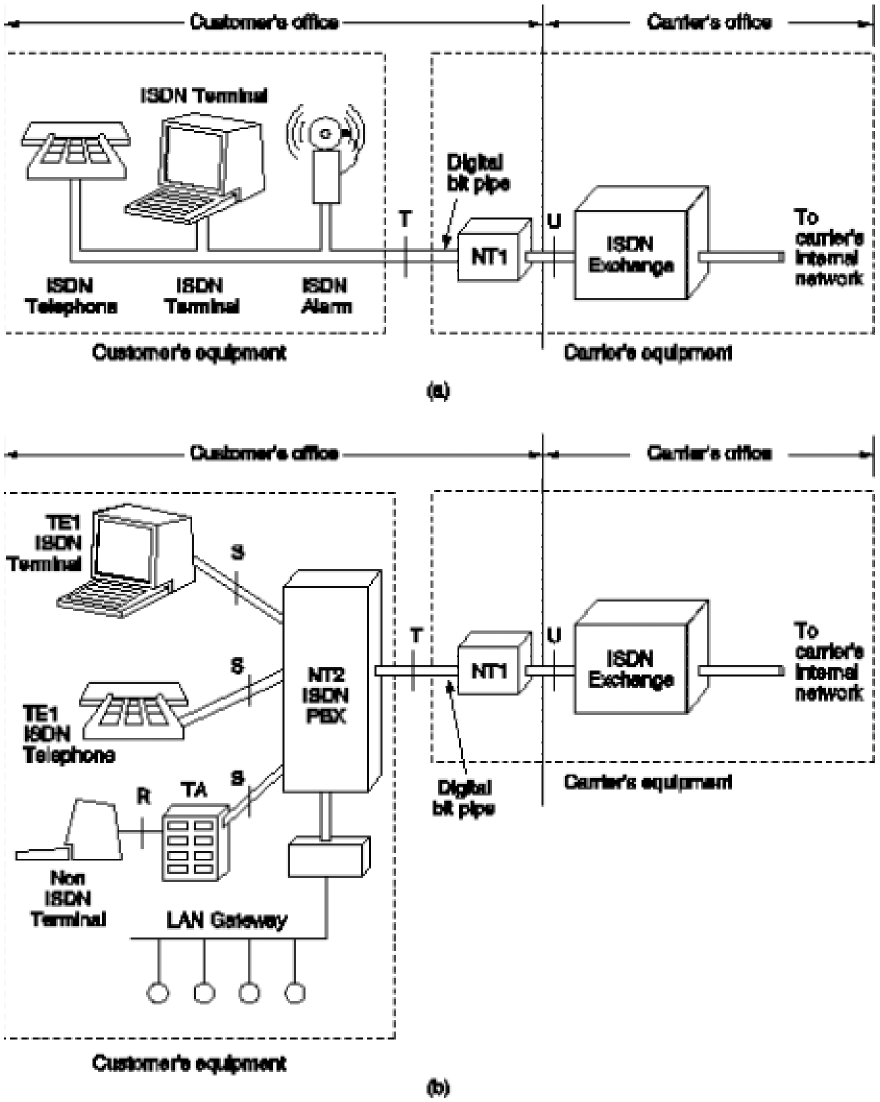


Fig. 2-41. (a) Example ISDN system for home use. (b) Example ISDN system with a PBX for use in large businesses.

For large businesses, the model of Fig. 2-41(b) is used. There is a device NT2 called PBX (Private Branch eXchange - conceptually the same as an ISDN switch) there connected to NT1 and providing the interface for ISDN devices.

CCITT defined four reference points (Fig. 2-41):

- U reference point = connection between the ISDN exchange and NT1,
- T reference point = connector on NT1 to the customer,
- S reference point = interface between the ISDN PBX and the ISDN terminal,
- R reference point = the connection between the terminal adapter and non-ISDN terminal.

2.5.3. The ISDN Interface

The ISDN bit pipe supports multiple channels interleaved by time division multiplexing. Several channel types have been standardized:

- A - 4 kHz analog telephone channel

- B - 64 kbps digital PCM channel for voice or data
- C - 8 or 16 kbps digital channel for out-of-band signaling
- D - 16 kbps digital channel for out-of-band signaling
- E - 64 kbps digital channel for internal ISDN signaling
- H - 384, 1536, or 1920 kbps digital channel.

It is not allowed to make arbitrary combination of channels on the digital pipe. Three combinations have been standardized so far:

- Basic rate: 2B + 1D. It should be viewed as a replacement for POTS (Plain Old Telephone Service). Each of the 64 kbps B channels can handle a single PCM voice channel with 8 bits samples made 8000 times per second. D channel is for signaling (i.e., to inform the local ISDN exchange of the address of the destination). The separate channel for signaling results in a significantly faster setup time.
- Primary rate: 23B + 1D (US and Japan) or 30B + 1D (Europe). It is intended for use at the T reference point for businesses with a PBX.
- Hybrid: 1A + 1C

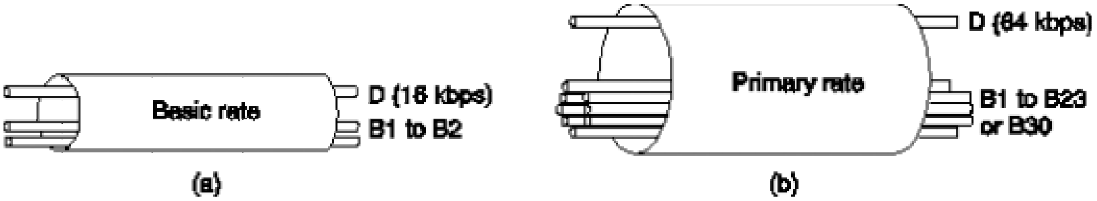


Fig. 2-42. (a) Basic rate digital pipe. (b) Primary rate digital pipe.

Because ISDN is so focused on 64 kbps channels, it is referred to as N-ISDN (Narrowband ISDN), in contrast to broadband ISDN (ATM).

2.5.4. Perspective on N-ISDN

N-ISDN was an attempt to replace the analog telephone system with a digital one. Unfortunately, the standardization process was too long and regarding to the technology progress in this area, once the standard was finally agreed, it was obsolete.

N-ISDN basic rate is too low so for home as for business today. N-ISDN may be partly saved, but by an unexpected application: Internet access. Various companies now sell ISDN adapters that combine the 2B + D channels into a single 144 kbps digital channel. Many Internet providers also support these adapters. So the people can access Internet over a 144 kbps digital link, instead of a 28.8 kbps analog modem link and for affordable price that may be a niche for N-ISDN for the next few years.

2.6. Broadband ISDN and ATM

When CCITT found that the N-ISDN was not going to solve the actual communication problems, it tried to think of a new service. The result was broadband ISDN (B-ISDN), basically a digital virtual circuit for moving fixed-sized packets (cells) at 155 Mbps.

Broadband ISDN is based on ATM technology that is fundamentally a packet-switching technology.

2.6.1. Virtual Circuits versus Circuit Switching

The basic broadband ISDN service is a compromise between pure circuit switching and pure packet switching. The actual service offered is connection oriented, but it is implemented internally with packet switching. Two kinds of connections are offered:

- Permanent virtual circuits - ordered by customers at carriers, remain in place for long time.
- Switched virtual circuits - set up dynamically like telephone calls.

The advantage of permanent over a switched virtual circuit is that there is no setup time, packets along permanent circuit can move instantly. For some applications, such as credit card verification, saving a few seconds on each transaction may be worth the cost of the permanent circuit.

In a virtual circuit network, like ATM, when a circuit is established, what really happens is that route is chosen from source to destination, and all the switches (i.e., routers) along the way make table entries so that they can route any packet on that virtual circuit (Fig. 2-43). When a packet comes along, the switch inspects the packet header to find out which virtual circuit it belongs to. Then it looks up that virtual circuit in its tables to determine which communication line to send on.

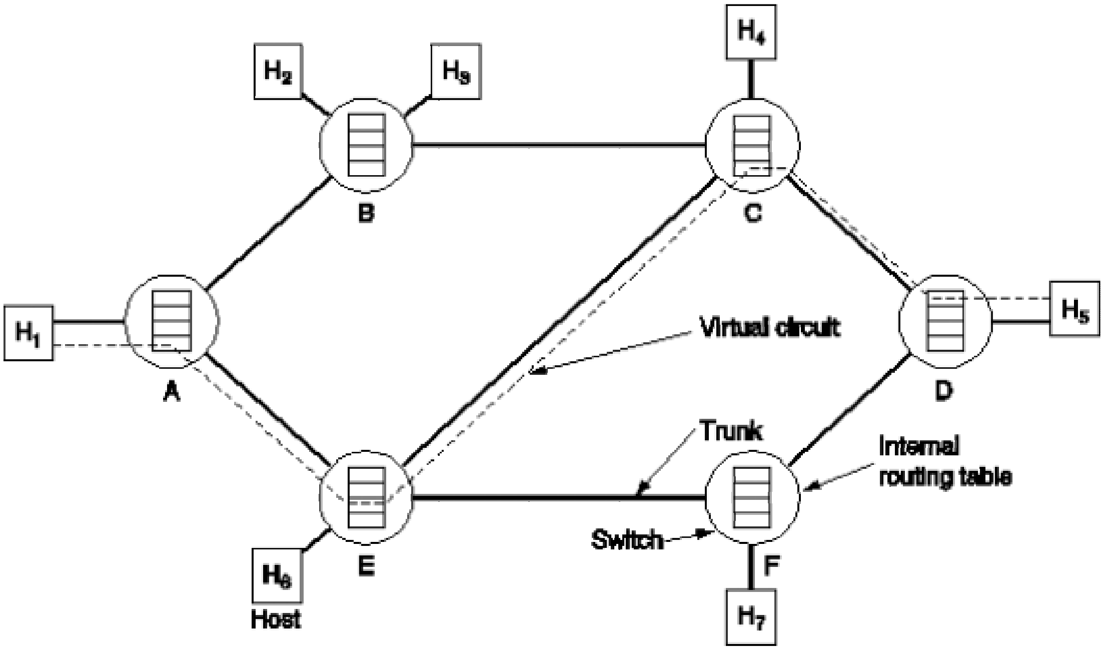


Fig. 2-43. The dotted line shows a virtual circuit. It is simply defined by table entries inside the switches.

2.6.2. Transmission in ATM Networks

ATM stands for Asynchronous Transfer Mode. This mode can be contrasted with the synchronous T1 carrier (Fig. 2-44).

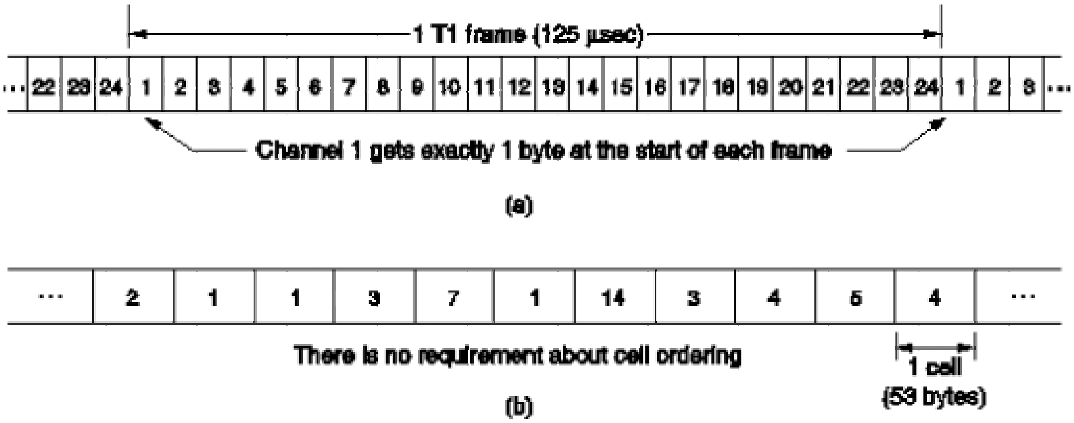


Fig. 2-44 (a) Synchronous transmission mode. (b) Asynchronous transmission mode.

T1: frames are generated precisely every 125(sec. This rate is governed by a master clock. Slot k of each frame contains 1 byte of data from the same source.

ATM: has no requirements that cells rigidly alternate among the various sources. Cells arrive randomly from different sources. The stream of cells need not be continuous. Gaps between the data are filled by special idle cells.

ATM does not standardize the format for transmitting cells. Cells are allowed to be sent individually, or they can be encased in a carrier such as T1, T3, SONET, or FDDI. For these examples, standards exist telling how cells are packed into the frames these systems provides.

In the original ATM standard, the primary rate was 155.52 Mbps, with an additional rate at four time that speed (622.08 Mbps). These rates were chosen to be compatible with SONET. ATM over T3 (44.736 Mbps) and FDDI (100 Mbps) is also foreseen.

The transmission medium for ATM is normally fiber optics, but for runs under 100 m, coax or category 5 twisted pair are also acceptable. Each link goes between a computer and an ATM switch, or between two ATM switches. So, all ATM links are point-to-point. Each link is unidirectional. For full-duplex operation, two parallel links are needed.

2.6.3. ATM Switches

An ATM cell switch (Fig. 2-45) has some number of input lines and some number of output lines (the both numbers are usually the same). ATM switches are generally synchronous in the sense of during a cycle, one cell is taken from each input line, passed into the internal switching fabric, and eventually transmitted on the appropriate output line.

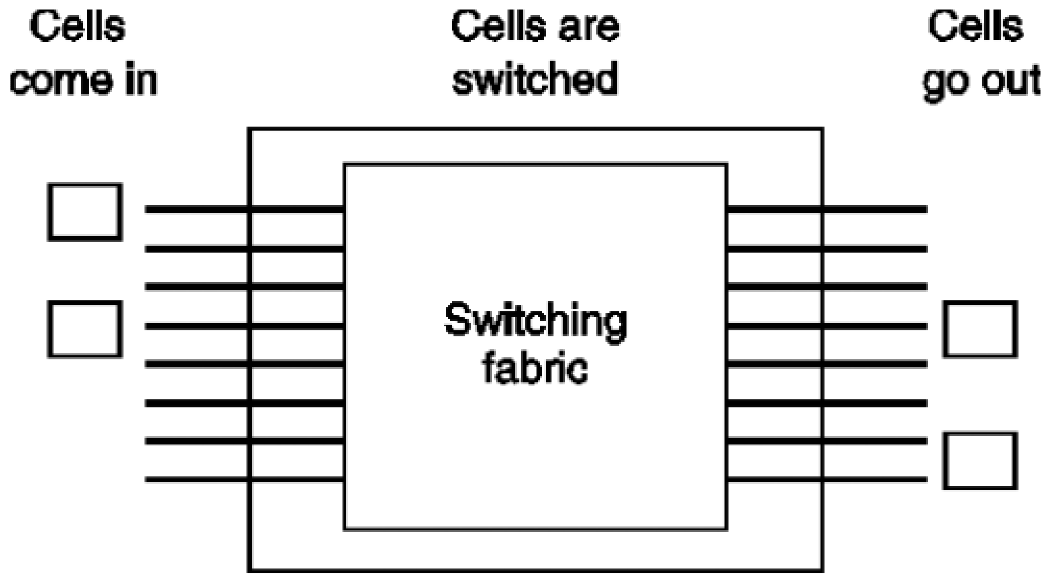


Fig. 2-45. A generic ATM switch.

Switches may be pipelined, that is, it may take several cycles before an incoming cell appears on its output line. Cells actually arrive on the input lines asynchronously, so there is a master clock that marks the beginning of a cycle. Any cell fully arrived when the clock ticks is eligible for switching during that cycle. A cell not fully arrived has to wait until the next cycle.

Cells arrive at ATM speed, normally about 150 Mbps. This works out around 360000 cells/sec, the cycle time has to be about 2.7 (sec. A commercial switch might have from 16 to 1024 input lines. At 622 Mbps the cycle time has to be about 700 nsec. The fact that the cells are fixed length and short makes it possible to build such switches. With longer variable-length packets, high speed switching would be more complex, which is why ATM uses short fixed-length cells.

All ATM switches have two common goals:

- Switch all cells with as low a discard rate as possible - cells can be dropped just in emergencies, but the loss rate should be as small as possible.
- Never reorder the cells on a virtual circuit - this constraint makes switch design considerably more difficult, but is required by the ATM standard.

A problem is if the cells arriving at more input lines want to go to the same output port in the same cycle.

One solution is to provide a queue for each input line. If more cells conflict, one of them is chosen for delivery, and the rest are held for the next cycle (Fig. 2-46). The problem with input queuing is that when a cell has to be held up, it blocks the progress of all cells behind it, even if they could otherwise be switched (head-of-line blocking).

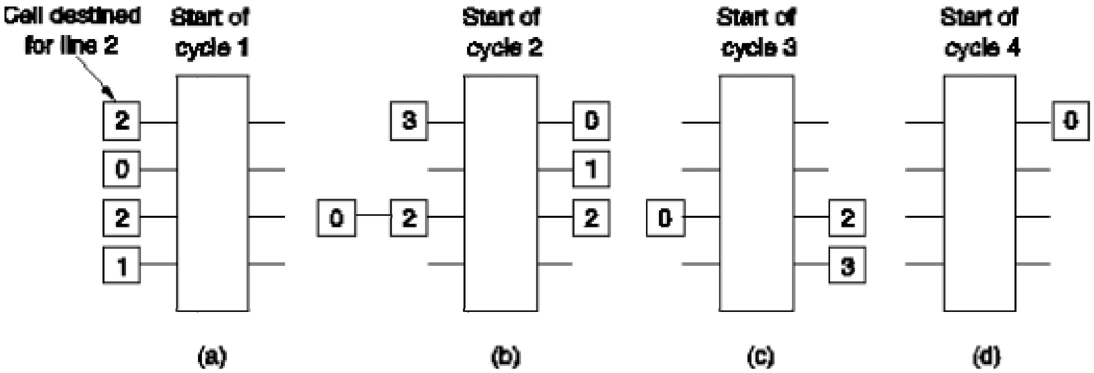


Fig. 2-46. Input queueing at an ATM switch.

An alternative design, one that does not exhibit head-of-line blocking, does the queuing on the output side (Fig. 2-47). In our example, it takes only three cycles, instead of four in the previous example, to switch all packets. Output queuing is generally more efficient than input queuing.

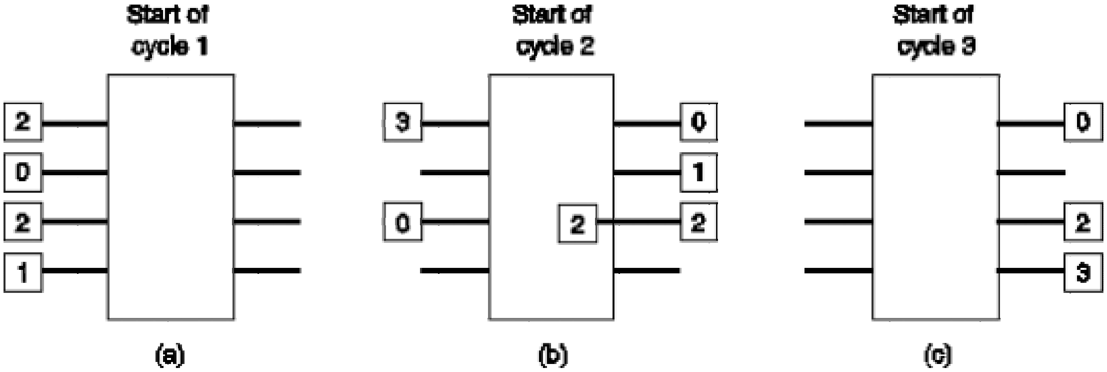


Fig. 2-47. Output queueing at an ATM switch.

2.6.4. The Knockout Switch

In Fig. 2-48, there is one ATM switch design, that uses output queuing, called knockout switch. Each input line is connected to a bus on which incoming cells are broadcasted in the cycle they arrive.

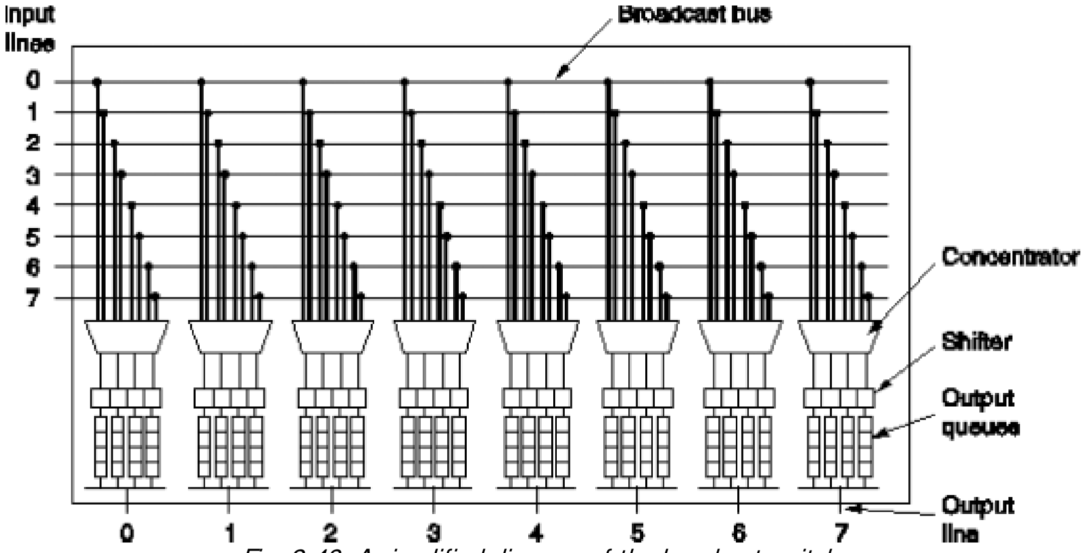


Fig. 2-48. A simplified diagram of the knockout switch.

For each arriving cell, hardware inspects the cell's header to find its virtual circuit information, looks up in the routing tables, and enables the correct crosspoint through which it gets to its output line. In case multiple cells want to go to the same output line a problem arises. The simplest way to solve such a problem is to buffer all cells at the output side. For switches with many inputs (say 1024) it is not reasonable to have a buffer for each output. In practice, reasonable optimization is to provide fewer output buffers, say n.

If more than n cells arrive in one cycle, the concentrator on each line selects out n cells for queuing, discarding the rest. It makes this selection using an elimination (knockout) tournament.

Conceptually, all the selected cells go into a single output queue (unless it is full, in which case cells are discarded). Because of timing reasons, the output queue is simulated by multiple queues. The selected cells go into a shifter, which then distributes them uniformly over n output queues using a token to keep track of which queue goes next, in order to maintain sequencing within each virtual circuit.

2.6.5. The Batcher-Banyan Switch

The problem with the knockout switch is that the number of crosspoints is quadratic in the number of lines. As with the crossbar switches for circuit switching, the solution is the space division switching requiring a multistage switch. This solution is called the Batcher-banyan switch.

In banyan switches, only one path exists from each input line to each output line (see Fig. 2-49(a) for 8 x 8 three-stage banyan switch). Routing is done by looking up the output line for each cell (based on virtual circuit information and tables). This 3 bit binary number is then put in front of the cell, as it will be used for routing through the switch.

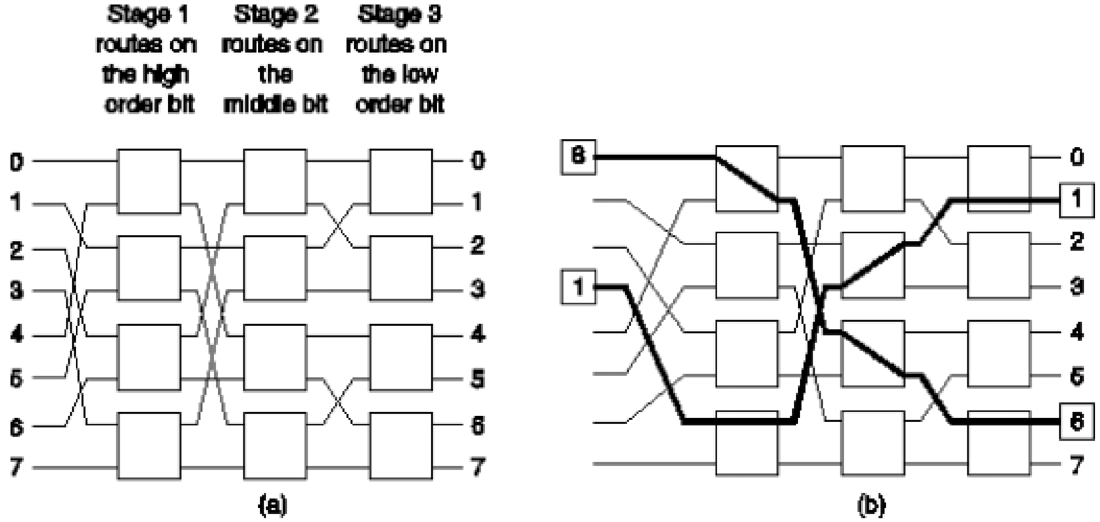

Stage 1 routes on the high order bit
Stage 2 routes on the middle bit
Stage 3 routes on the low order bit
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
7
(a)
6
1
0
1
2
3
4
5
6
7
(b)

Fig. 2-49. (a) A banyan switch with eight input lines and eight output lines.
(b) The routes that two cells take through the banyan switch.

Each of the 12 switching elements in the banyan switch has two inputs and 2 outputs. When a cell arrives at a switching element, 1 bit of the output line number is inspected, and based on that, the cell is routed either to port 0 (the upper one) or port 1 (the lower one). In the event of collision, one cell is routed and one is discarded.

A banyan switch parses the output line number from left to right (see Fig. 2-49(b) for an example).

Examples in Fig. 2-50 show that depending on input, the banyan switch can do a good job or a bad job of routing.

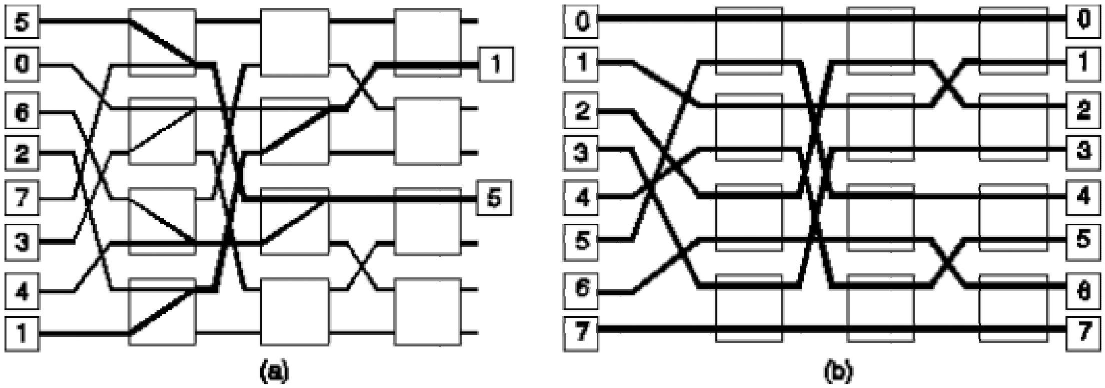


Fig. 2-50. (a) Cells colliding in a banyan switch. (b) Collision-free routing through a banyan switch.

The idea behind the Batcher-banyan switch is to put a switch in front of the banyan switch to permute the cells into a configuration that the banyan switch can handle without loss. For example, if the incoming cells are sorted by destination and presented on input lines 0, 2, 4, 6, 1, 3, 5, and 7, in that order as far as necessary (depending of how many cells there are), then the banyan switch does not lose cells.

To sort the incoming cells we can use a Batcher switch built up of 2 x 2 switching elements. When such a switching element receives two cells, it compares their output addresses numerically (thus no just 1 bit) and routes the higher one on the port in the direction of arrow, and the lower one on the other way. If there is only one cell, it goes to the port opposite the way the arrow is pointing (Fig. 2-51).

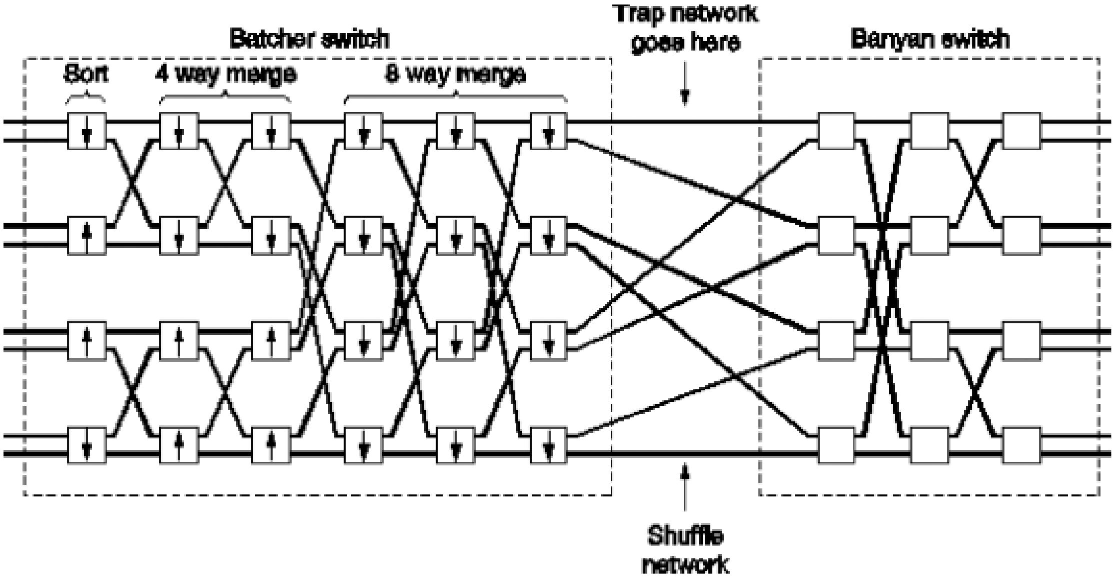


Fig. 2-51. The switching fabric for a Batcher-banyan switch.

After exiting the Batcher switch, the cells undergo a shuffle and are then inserted into a banyan switch (see Fig. 2-52 for an example).

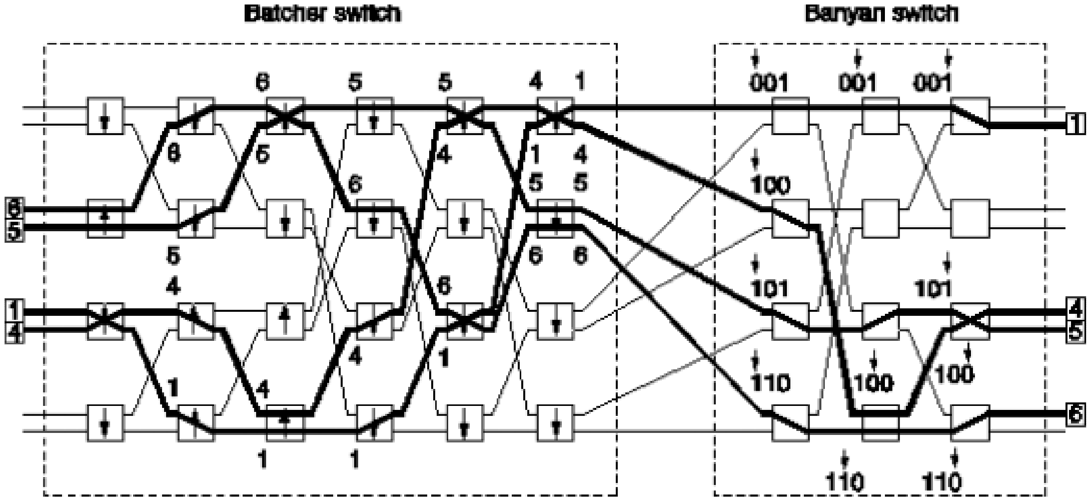


Fig. 2-52. An example with four cells using the Batcher-banyan switch.

In principle, the Batcher-banyan switch makes a fine ATM switch, but there are two complications: output line collision and multicasting. If two or more cells are aimed at the same output, the switch cannot handle them, so a kind of buffering has to be introduced. One way to solve this problem is by inserting a trap network between the Batcher switch and banyan switch that filters out duplicates and recirculates them for subsequent cycles, all the while maintaining the order of cells on a virtual circuit. Commercial switches also have to handle multicast.

2.7. Cellular Radio

The traditional telephone system (even when broadband ISDN is fully operating) will still not be able to satisfy people on the go. Consequently, there is increasing competition from systems that use wireless technologies for communication. They are already creating a huge market. Many companies in the computer, telephone, satellite, and other industries want a piece of action. The result is a chaotic market, with numerous overlapping and incompatible products and services, all rapidly changing.

2.7.1. Paging Systems

The first paging systems used loudspeakers within a single building. Nowadays, people who want to be paged wear small beepers, usually with tiny screens for displaying short incoming messages.

A person wanting to page a beeper wearer can call the beeper company and enter a security code, the beeper number, and the number the beeper wearer is to call (or another short message). The request is then broadcasted from a hilltop antenna (for local paging) or from a satellite (for long distance paging). When a beeper detects its unique number in the incoming radio stream, it beeps and displays the number to be called.

Most current paging systems are one-way systems, from a single computer out to a large number of receivers. There is no problem about who will speak next, and no contention among many competing users.

Paging systems require little bandwidth since each message requires only a single burst of about 30 bytes. At this rate, a 1 Mbps satellite channel can handle over 240000 pages per minute. The older paging systems run in the 150 - 174 MHz band, the modern ones in the 930 - 932 MHz band. (Fig. 2-53).

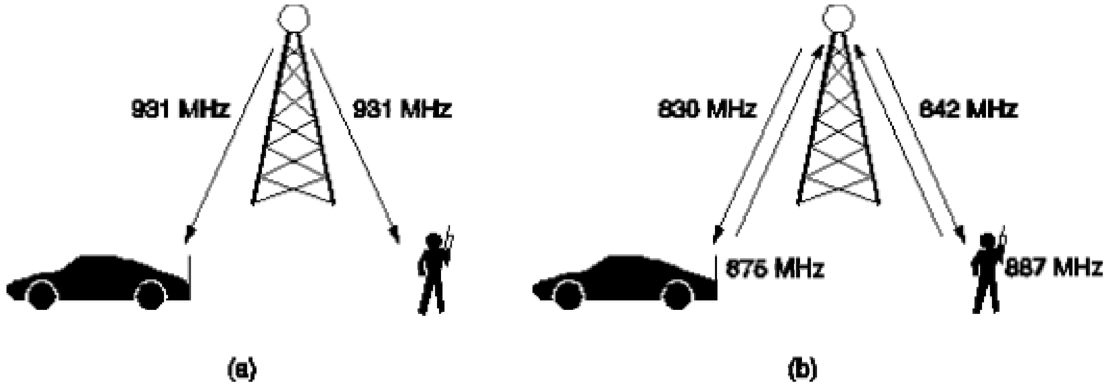


Fig. 2-53. (a) Paging systems are one way. (b) Mobile telephones are two way.

2.7.2. Cordless Telephones

A cordless telephone consists of two parts: a base station and a telephone. The base station has a standard phone jack and is connected by a wire to the telephone system. The telephone communicates with the base station by low-power radio. The range is typically 100 - 300 m.

Some of cheaper models of cordless telephones used a fixed frequency, selected at the factory. If, by accident, someone in the neighborhood of a user had the telephone with the same frequency, he could listen user's calls. More expensive models avoided this problem by allowing the user to select the transmission frequency.

The generations of cordless telephones:

- CT-1 in the US and CEPT-1 in Europe - entirely analog. They could cause interference with radios and television. Poor reception and lack of security.
- CT-2 - digital standard, originated in England. Each telephone had to be within a few hundred meters of its own base station. Useful around the house or office, useless in cars when walking around the town.
- CT3 or DECT - third generation introduced in 1992. This technology is beginning to approach cellular telephones.

2.7.3. Analog Cellular Telephones

Mobile radiotelephones were used sporadically for maritime and military communication during the early decades of the 20th century.

Push-to-talk systems (installed in several big cities in the late 1950s) had a single channel used for both sending and receiving. They used a large transmitter on top of a tall building. To talk, the user had to push a button that enabled the transmitter and disabled the receiver. The users could hear each other.

IMTS (Improved Mobile Telephone System - in the 1960s) also used a high-powered (200 watt) transmitter, on top of a hill, but it used different frequencies for sending and for receiving, so no push-to-talk button was necessary. IMTS supported 23 channels spread out from 150 MHz to 450 MHz. Due to the small number of channels, users often had to wait a long time before getting a dial tone. The adjacent systems had to be several hundred km apart. So the system was impractical due to limited capacity.

2.7.4. Advanced Mobile Phone System

AMPS (Advanced Mobile Phone System) was invented by Bell Labs, first installed in US in 1982. It is also used in England, where it is called TACS, and in Japan, where it is called MCS-L1.

In AMPS, a geographic region is divided up into cells, typically 10 to 20 km across, each using some set of frequencies. The key idea that gives AMPS far more capacity than all previous systems, is using relatively small cells, and reusing transmission frequencies in nearby (but not adjacent) cells. The idea of frequency reuse is illustrated in Fig. 2-54(a). The cells are normally roughly circular, but they are easier to model as hexagon. In Fig. 2-54(a), the cells are all the same size. They are grouped in units of seven cells. Each letter indicates a group of frequencies.

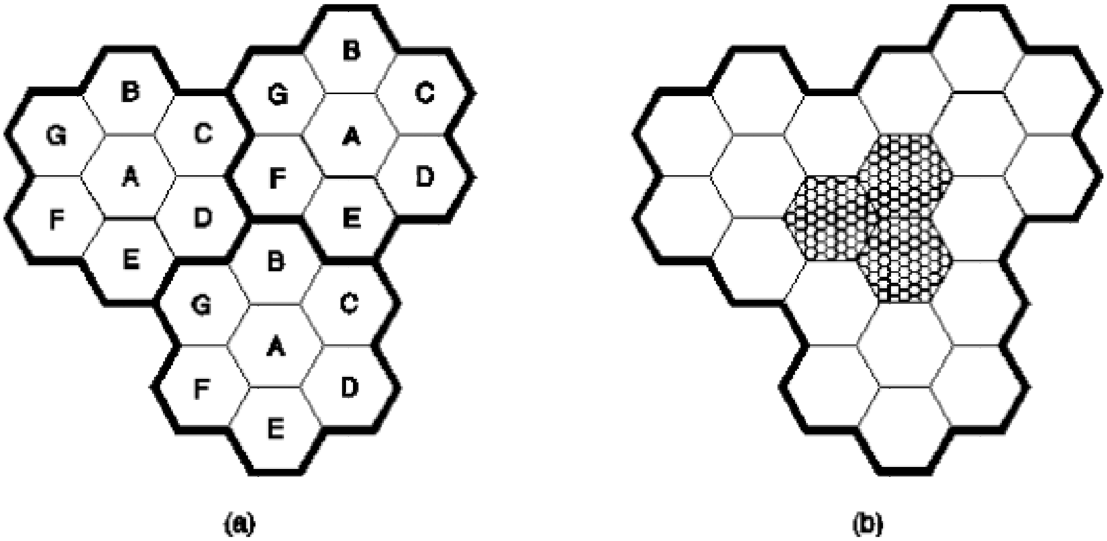


Fig. 2-54. (a) Frequencies are not reused in adjacent cells. (b) To add more users, smaler cells can be used.

In an area where the number of users has grown to the point where the system is overloaded, the power is reduced and the overloaded cells are split into smaller ones to permit more frequency reuse (Fig. 2-54(b)).

At the center of each cell, there is a *base station* to which all the telephones in the cell transmit. The base station consists of a computer and transmitter/receiver connected to an antenna. The base stations are connected to *MTSO* (Mobile Telephone Switching Office). In larger areas, several MTSOs may be needed, all of which are connected to a second-level MTSO, and so on. The MTSO system is connected to at least one telephone system end office. The MTSOs communicate with the base stations, each other, and the PSTN using a packet switching network.

At any instant, each mobile telephone is logically in one specific cell and under the control of that cell's base station. When a mobile telephone leaves a cell, its ownership is transferred to the cell getting the strongest signal from it. If a call is in progress, it will be asked to switch to a new channel. This process is called handoff and takes about 300 msec. The channel assignment is done by MTSO.

2.7.5. Channels

The AMPS system uses 832 full-duplex channels, each consisting of a pair of simplex channels. There are 832 simplex transmission channels from 824 to 849 MHz, and 832 simplex receive channels from 869 to 894 MHz. Each of these simplex channels is 30 kHz wide. AMPS uses FDM to separate the channels.

In the 800 MHz band, radio waves travel in straight lines. They are absorbed by trees and plants and bounce off the ground and buildings. This may lead to an echo effects or signal distortion.

The 832 channels are divided into four categories:

1. Control (base to mobile) to manage the system. 21 channels are reserved for control, and these are wired into a PROM in each telephone.
2. Paging (base to mobile) to alert mobile users to call for them.
3. Access (bi-directional) for call setup and channel assignment.
4. Data (bi-directional) for voice, fax, and data.

Since the same frequencies cannot be reused in nearby cells, the actual number of voice channels per cell is much smaller than 832, typically about 45.

2.7.6. Call Management

Each mobile phone in AMPS has a 32 bit serial number and 10 digit telephone number in its PROM. When a phone is switched on, it scans a preprogrammed list of 21 control channels to find the most powerful signal. From the control channel, it learns the number of paging and access channels.

The phone then broadcasts its serial number and telephone number several times. When the base station hears the announcement, it tells the MTSO, which records the existence of its new customer and also inform the customer’s home MTSO of his current location. During the normal operation, the mobile telephone reregisters about once every 15 minutes.

To make a call, the user enters the number to be called, and hits the send button. The phone sends the number and its own identity on the access channel. When the base station gets the request, it informs the MTSO. The MTSO looks for idle channel for the call. If one is found, the channel number is sent back on the control channel. The mobile phone then automatically switches to the selected voice channels and waits until the called party picks up the phone.

As for incoming calls, all idle phones continuously listen to the paging channel to detect messages directed at them. When a call is placed to a mobile phone, a packet is sent to the callee’s home MTSO to find out where it is. A packet is then sent to the base station in its current cell, which then sends a broadcast on the paging channel of the form: "Unit 14, are you here?" The called phone then responds with "Yes" on the control channel. The base then says something like: "Unit 14, call for you on channel 3." The called phone switches to channel 3 and starts making ringing sounds.

2.7.7. Security Issues

Analog cellular phones are totally insecure. Anyone with an all-band radio receiver (scanner) can tune in and hear everything going in a cell.

Another major problem is theft of air time, again based on the possibility of monitoring the transmitted information.

Some of these problems could be solved by encryption, but then the police could not easily perform "wiretaps" on wireless criminals.

2.7.8. Digital Cellular Telephones

First generation cellular systems were analog. The second generation is digital. In Europe, an agreement on a common digital system, call *GSM* (Global System for Mobile communication) was achieved.

GSM operates in a new frequency band (1.8 GHz) and uses both FDM and TDM. The available spectrum is broken up into 50 200 kHz bands. With each band TDM is used to multiplex multiple users.

Some GSM telephones use smart cards (credit card sized devices containing a CPU). The serial number and telephone number are contained there, not in telephone, making for better security. Encryption is also used.

2.7.9. Personal Communication Services

Everyone would like to have a small cordless phone that works around the house and also anywhere in the world. Such a system is currently under vigorous development. In the US it is called PCS (Personal Communication Services), everywhere else it is called PCN (Personal Communication Network).

PCS will use cellar technology, but with microcells, perhaps 50 - 100 m wide. This allows very low power (1/4 watt), which makes it possible to build very small, light phones. On the other hand, it requires many more cells. The small base stations in these cells are sometimes called telepoints.

2.8. Communication Satellites

In the 1950s and early 1960s people tried to set up communication systems by bouncing signals off metalized balloons. Unfortunately, the received signals were too week to be of any practical use. The US Navy built an operational system for ship-to-shore communication by bouncing signal off the moon.

Further progress in the celestial communication came with the first communication satellite launched in 1962. The artificial satellites can amplify the signal before sending them back and can be thought as a big microwave repeaters in the sky.

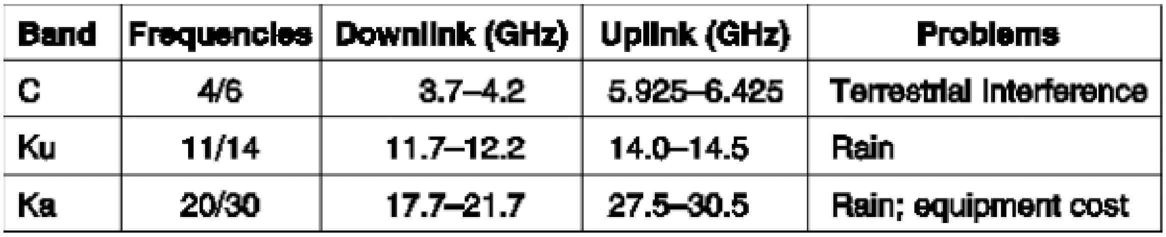
Communication satellites contain several transpoders, each of which listens to some portion of the spectrum, amplifies the incoming signal, and then rebroadcasts it at another frequency, to avoid interference with the incoming signal. The downward beams can be broad, covering a substantial fraction of the earth’s surface, or narrow, covering an area only hundreds of kilometers in diameter.

2.8.1. Geosynchronous Satellites

According to Kepler’s law, the orbital period of a satellite varies as the orbital radius to the 3/2 power. Near the surface of the earth, the period is about 90 minutes that use useless for communication satellites. However, at an altitude approximately 36000 km above the equator, the satellite period is 24 hours, so it revolves at the same rate as the earth under it. It is very desirable for communication purposes.

With current technology, it is in general unwise to have satellites spaced much closer than 2 degrees in the 360 degree equatorial plane, to avoid interference. So we have just 180 slots for geosynchronous satellites. Fortunately, satellites using different parts of spectrum do not compete, so each of the 180 possible satellites could have several data streams going up and down simultaneously. Alternatively, two or more satellites could occupy one orbit slot if they operate at different frequencies.

There have been international agreement about who may use which orbit slots and frequencies. The main commercial bands are the following (Fig. 2-55):



Band	Frequencies	Downlink (GHz)	Uplink (GHz)	Problems
C	4/6	3.7–4.2	5.925–6.425	Terrestrial interference
Ku	11/14	11.7–12.2	14.0–14.5	Rain
Ka	20/30	17.7–21.7	27.5–30.5	Rain; equipment cost

Fig. 2-55. The principal satellite bands.

1. C band - the frequency ranges assigned to this band are already overcrowded because they are also used by the common carriers for terrestrial microwave links.
2. Ku band - this band is not (yet) congested, and at these frequencies satellites can be spaces as close as 1 degree. But another problem exists: absorption by rain. This problem can be circumvented by using several widely separated ground stations.
3. Ka band - the problem with this band is still expensive equipment.

In addition to these commercial bands, many government and military bands also exist.

A typical satellite has 12-20 transpoders, each with 36-50 MHz bandwidth (e.g., a 50 Mbps transpoder can handle 800 64 kbps digital voice channels).

The first satellites had a single spatial beam that illuminated the entire earth. Nowadays, each downward beam can be focused on a small geographical area, typically elliptically shaped, and as small as a few hundreds km in diameter (so called spot beams).

A new development in the communication satellite world is the development of low-cost microstations, sometimes called VSATs (Very Small Aperture Terminals). These tiny terminals can put out only small power (1 watt) and the communication among them is ensured by using a special ground station, the hub, with a large antenna and amplifier (Fig. 2-56). The trade off is a longer delay.

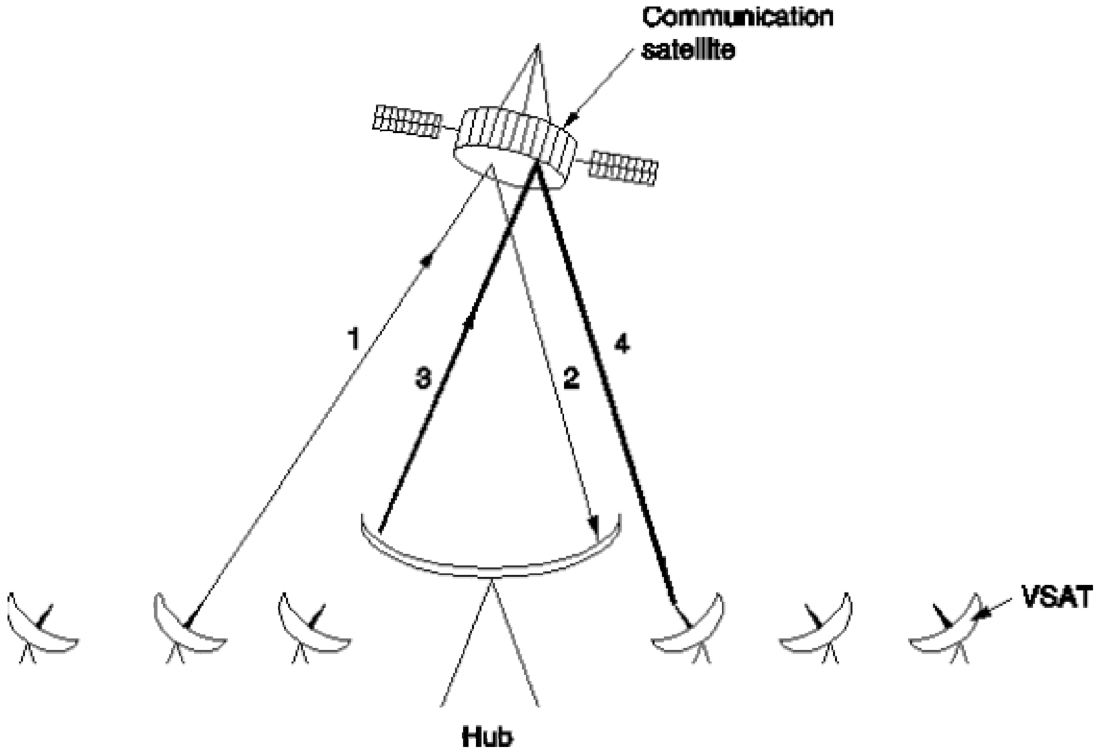


Fig. 2-56. VSATs using a hub.

Differences between communication satellites communication and terrestrial microwave links communication:

- Average end-to-end transmit time 270 msec at satellites (540 at VSATs), much longer than at terrestrial communication).
- Satellites are inherently broadcast media, the satellite broadcasting is much cheaper that terrestrial one.
- If security is required, encryption must be used at satellite transmission.
- At satellite communication, the cost for transmitting of a message is independent of the distance of the source and destination. A call across the ocean costs no more to service than a call across the street.
- Satellites have excellent error rates.

2.8.2. Low-Orbit Satellites

For the first 30 years of satellite era, low-orbit satellites were rarely used for communication because they zip into and out of view so quickly. In 1990, Motorola started a new activity called Iridium project, aimed to communication based on low orbit satellites.

The basic goal of Iridium is to provide worldwide telecommunication service using hand-held devices that communicate directly with Iridium satellites. This service competes with PCS/PCN activities.

The project uses ideas from cellular radio, but with moving cells. The satellites beams scan the earth as the satellites move. The handover techniques used in cellular radio are applicable also in this case.

The satellites (66 in total) are to be positioned at an altitude of 750 km, in circular polar orbits (Fig 2-57(a)). With 6 satellite necklaces, the entire earth would be covered (Fig. 2-57(b)).

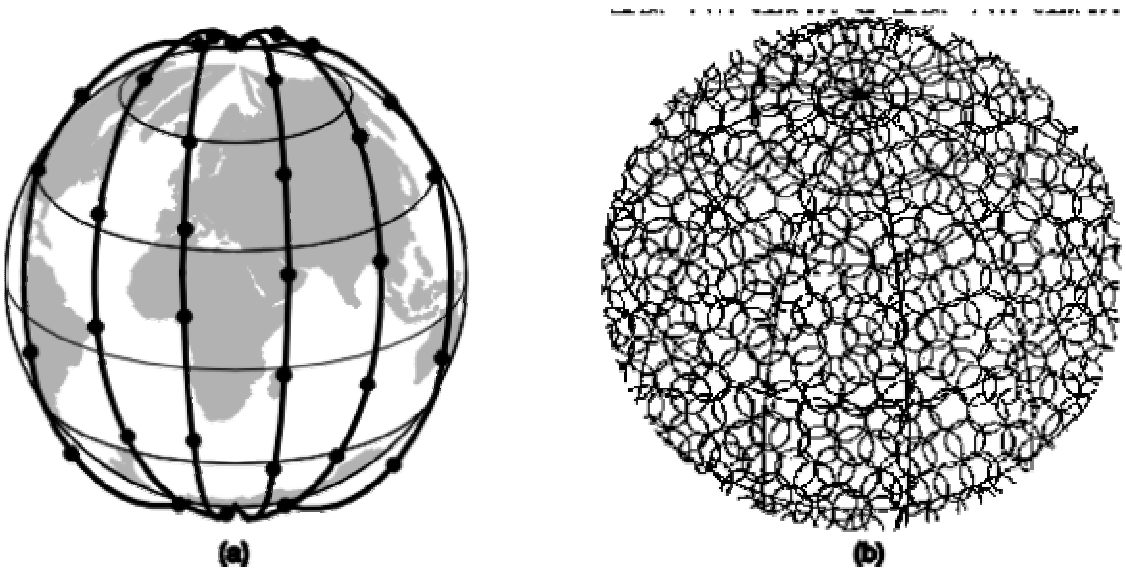


Fig. 2-57. (a) The Iridium satellites from six necklaces around the earth. (b) 1628 moving cells cover the earth.

2.8.3. Satellites versus Fiber

A comparison between satellite communication and terrestrial communication is instructive.

20 years ago, it could seem that the future of communication is in communication satellites. Telephone system had changed little in the past 100 years and showed no signs of changing in the next 10 years. This glacial movement was caused in no small part by the regulatory environment in which telephone companies were expected to provide good voice services at reasonable prices (which they did), and in return got a guaranteed profit on their investments. For people with data to transmit, 1200 bps modems were available. That was pretty much all there was.

The introduction of competition in 1984 in the US and somewhat later in Europe change all that radically. Telephone companies began replacing their long-haul networks with fiber and introduced high bandwidth services. They also stopped of charging artificially high prices to long-distance users to subsidize local service. In this situation, terrestrial fiber connections looked like the long-term winner.

Nevertheless, communication satellites have some major niche market that fiber does not address. They are:

- The high bandwidth of fiber is not available to most users. Satellites can easily provide high bandwidth to single users.
- Terrestrial fiber optic link cannot provide mobile communication. The satellites can.
- Systems based on satellites communication can more easily provide broadcasting.
- Satellite communication is much easier to install in places with hostile terrain or a poorly developed terrestrial infrastructure.
- Satellite communication is convenient where obtaining the right for laying cables is difficult or very expensive.
- Satellite communication is preferable when rapid deployment is critical (military communication systems in time of war).

In short, it looks like the mainstream communication of the future will be terrestrial fiber optics combined with cellular radio, but for some specialized uses, satellites are better. However, economics can influence all of this. Terrestrial and satellite communication will compete on price. If advances of technology radically reduce the cost of deploying a satellite, or low-orbit satellites catch on, it is not certain that fiber will win in all markets.

3. The Network Layer

3.1. Internetworking

In this section we will take a look at the issues that arise when two or more networks are together to form an *internet*.

Many network types are in operation today and it seems that this trend will continue also in the future. An example of different networks interconnection is in Fig. 5-33.

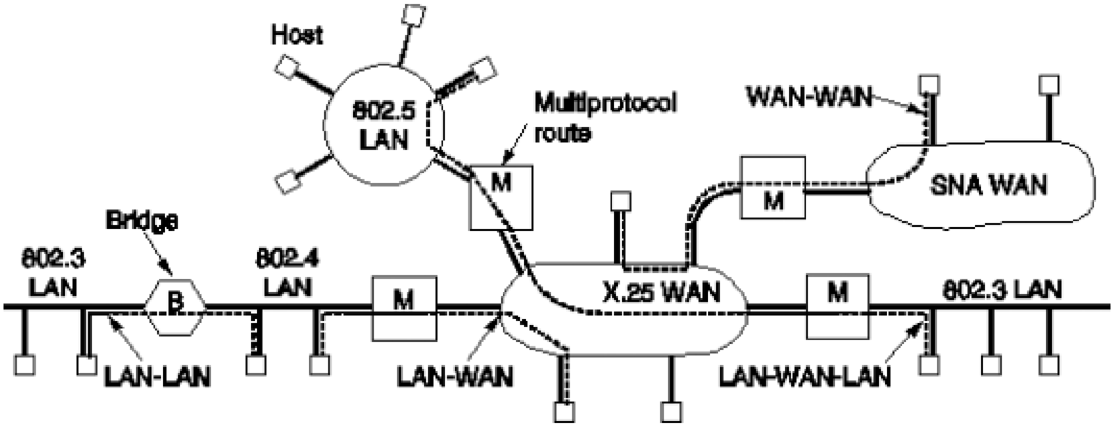


Fig. 5-33. Network interconnection.

At the junction between two networks that have to be interconnected, a "black box" for handling the necessary conversions as packets move from one network to the other has to be inserted. The name used for the black box depends on the layer that does the work. Although there is not much agreement on terminology in this area, usually the following names are used:

Layer 1: *Repeaters* copy individual bits between cable segments. They are used just to amplify or regenerate weak signals.

Layer 2: *Bridges* store and forward data link frames between LANs. A bridge accepts an entire frame and passes it up to the data link layer where the checksum is verified. Then the frame is sent down to the physical layer for forwarding on a different network.

Layer 3: *Multiprotocol routers* forward packets between dissimilar networks. They operate at the level of network layer.

Layer 4: *Transport gateways* connect byte stream in the transport layer.

Above 4: *Application gateways* allow interworking above layer 4. As an example we can take mail gateways.

For convenience, we will sometimes use the term "gateway" to mean any device that connects two or more dissimilar networks.

A gateway can be ripped apart in the middle and the two parts connected by a wire. Each of the halves is called a *half-gateway* (Fig. 5-34).

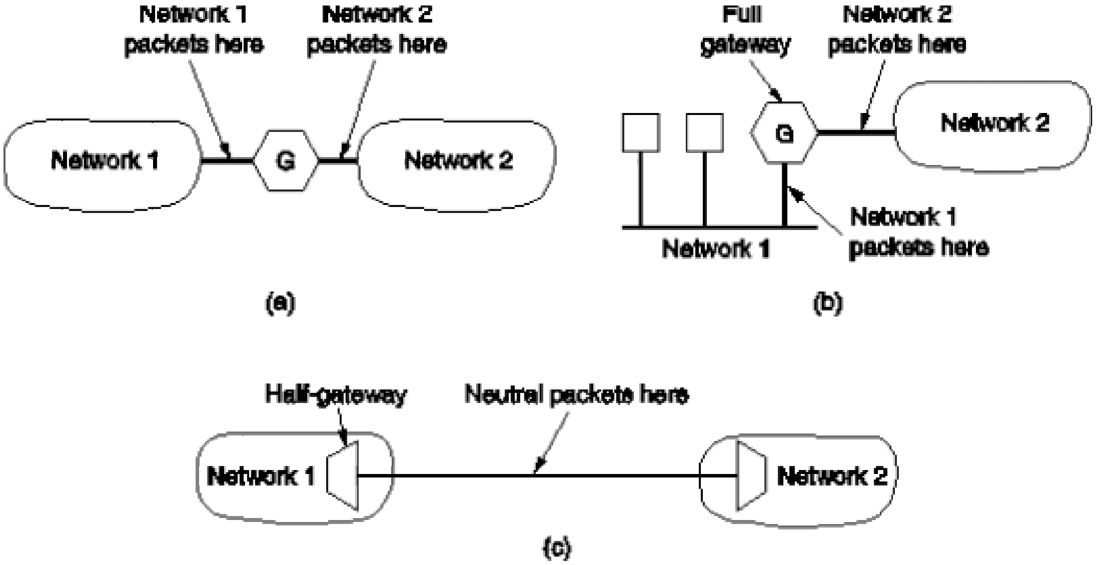


Fig. 5-34. (a) A full gateway between two WANs. (b) A full gateway between a LAN and a WAN. (c) Two half-gateways.

The situation in practice can be a little different than in theory. Many devices on the market combine bridge and router functionality and moreover some of them are sold under a wrong label.

3.1.1. How Networks Differ

Networks can differ in many ways.

3.1.2. Concatenated Virtual Circuits

Two *styles of internetworking* are common:

- a connection-oriented concatenation of virtual circuit subnets,
- a datagram internet style.

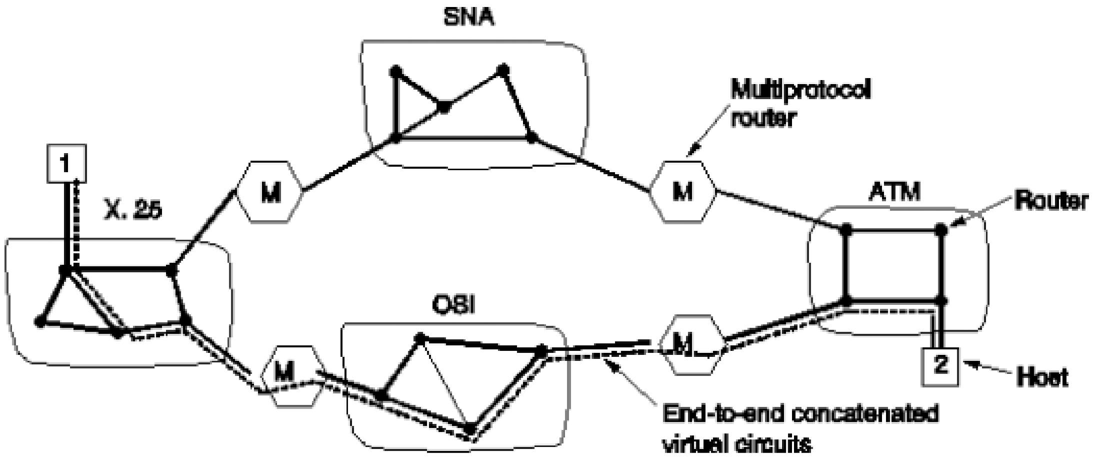


Fig. 5-36. Internetworking using concatenated virtual circuits.

In the concatenated virtual circuit model (Fig. 5-36) a connection to a host in a distant network is set up in a way similar to the way connections are normally established. The virtual circuit consists of concatenated virtual circuits between the routers or gateways along the way from the source node to the destination node. Each gateway maintains tables telling which virtual circuits pass through it, where they are to be routed, and what the new virtual circuit number is.

3.1.3. Connectionless Internetworking

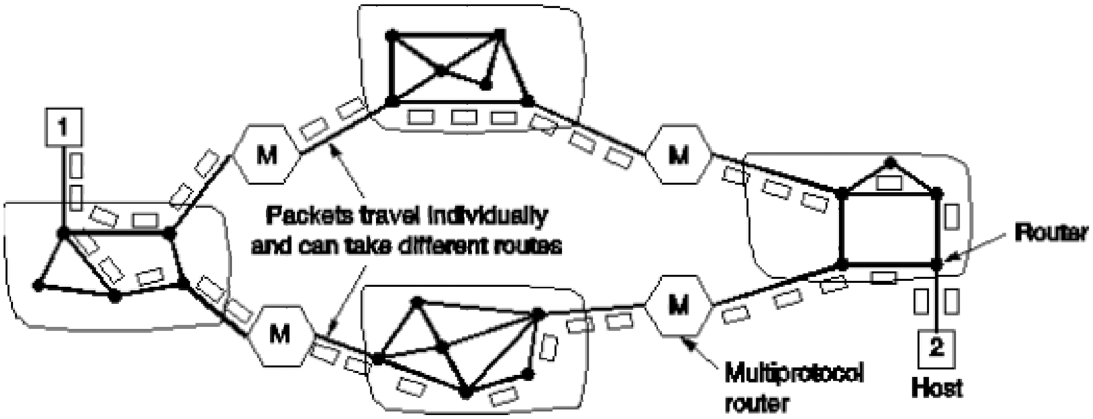


Fig. 5-37. A connectionless internet.

The alternative internetwork model is the *datagram model* (Fig. 5-37). In this model, the network layer offers to the transport layer just the ability to inject datagram into the subnet and hope it will get to the destination. Not all packet from a source to the same destination traverse the same sequence of gateways. A routing decision is made separately for each packet possibly depending on the traffic at the moment the packet is sent.

Complications leading often to insurmountable problems with the internetworking arise when:

- each network has its own network layer protocol,
- each network has its own addressing.

3.1.4. Tunneling

Handling the general case of making two different networks interwork is exceedingly difficult. However, in the special case, when the source and destination hosts are on the same type of network, but there is a different network in between, the situation is manageable.

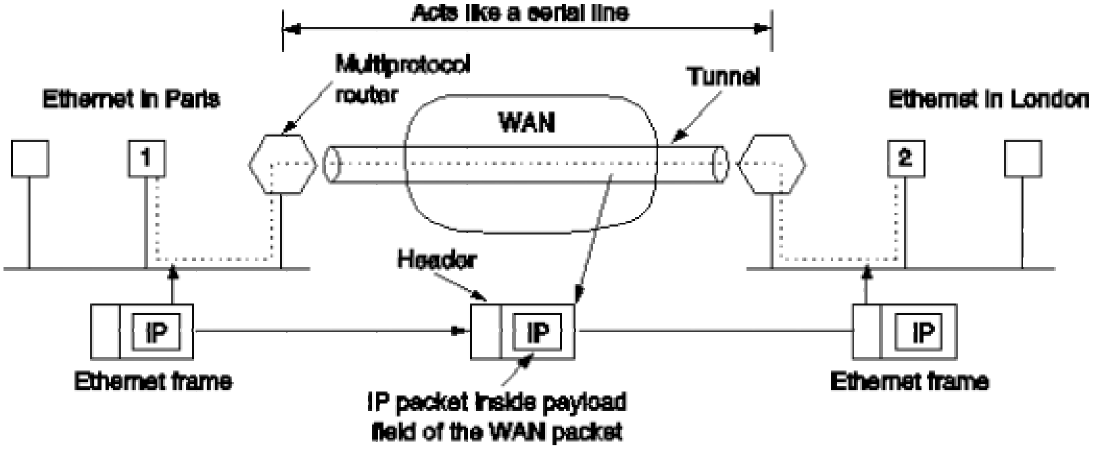


Fig. 5-38. Tunneling a packet from Paris to London.

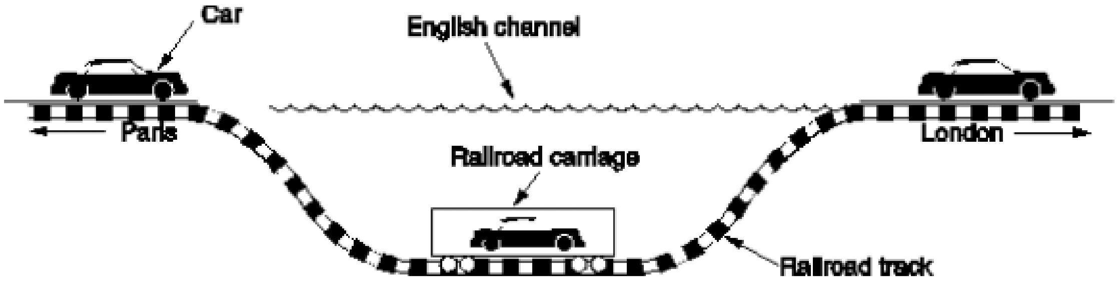


Fig. 5-39. Tunneling a car from France to England.

The solution to this problem is a technique called tunneling. In the Fig. 5-38, to send an IP packet to host 2, host 1 constructs the packet containing the IP address of host 2, inserts it into an Ethernet frame addressed to the Paris multiprotocol router, and puts it on the Ethernet. The multiprotocol router removes the IP packet, inserts it in the payload field of the WAN network layer packet, and addresses the later to the WAN address of the London multiprotocol router. When it gets there, the London router removes the IP packet and sends it to host 2 inside an Ethernet frame. So the WAN can be seen as a big tunnel extending from one multiprotocol router to the other.

3.1.5. Internetwork routing

Routing through an internetwork is similar to routing within a single subnet, but with some added complications. In the situation depicted in Fig. 5-40, every multiprotocol router (or gateway) can directly access (i.e. send packets to) every other router connected to any network to which it is connected. This leads to the graph model of the situation displayed in the b-part of the figure.

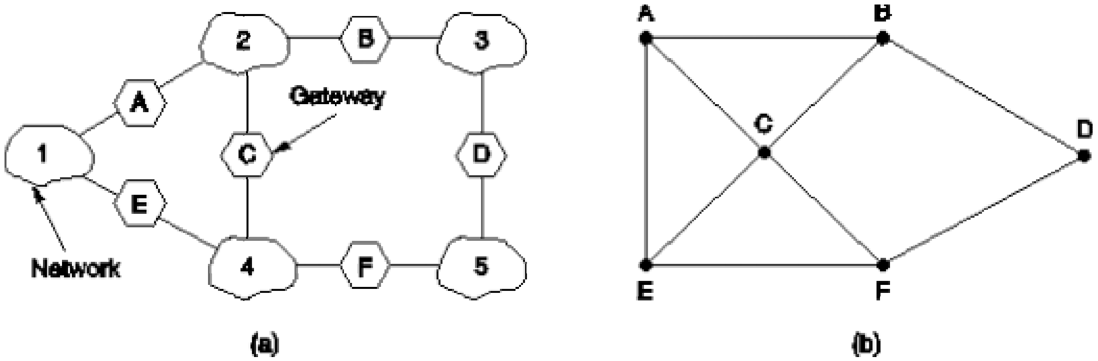


Fig. 5-40. (a) An internetwork. (b) A graph of the internetwork.

The typical *routing process* in such an internetwork looks as follows: an internet packet starts on its LAN addressed to the local multiprotocol router (in the MAC layer header). After it gets there, the network layer code decides which multiprotocol router to forward the packet to using its own routing tables. If that router can be reached using the packet's native network protocol, it is forwarded there directly. Otherwise it is tunneled there, encapsulated in the protocol required by the intervening network. This process is repeated until the packet reaches the destination network.

In the example above, two level routing algorithm has been applied: within each network an interior gateway protocol is used, but between the networks, an exterior gateway protocol is used. In fact, since each network is independent, they may all use different algorithms. Because each network in an internetwork is independent of all others, it is often referred to as Autonomous System (AS).

Internetwork routing often requires crossing international boundaries, where various laws come into play. This fact may insert different nontechnical elements and influences into the process of networking (e.g., by Canadian law, data traffic originating in Canada and ending in Canada may not leave the country. As a consequence, in some cases, non-optimal paths must be taken to deliver data through the network).

3.1.6. Fragmentation

Each network imposes some maximum size on its packets. These limits have various causes, among them:

1. Hardware (e.g., the width of a TDM transmission slot).
2. Operating system (e.g., all buffers are 512 bytes).
3. Protocols (e.g., the number of bits in the packet length field).
4. Compliance with some (inter)national standards.
5. Desire to reduce error induced retransmissions to some level.
6. Desire to prevent one packet from occupying the channel too long.

Maximum payload range from 48 bytes (ATM cells) to 65515 bytes (IP packets), although the payload size in higher layers is often larger.

A problem appears when a large packet wants to travel through a network whose maximum packet size is too small. The only solution to the problem is to allow gateways to break packets into fragments, sending each fragment as a separate internet packet. But then a new problem arises: how to put the fragments back together again.

Two opposing strategies exists for *recombining the fragments* back into the original packet:

- The fragments are recombined in the next gateway - so the fragmentation is made transparent to any subsequent network (Fig. 5-41(a)).
- Once a packet has been fragmented, each fragment is treated as though it were an original packet. Recombination occurs only at the destination host (Fig. 5-41(B)).

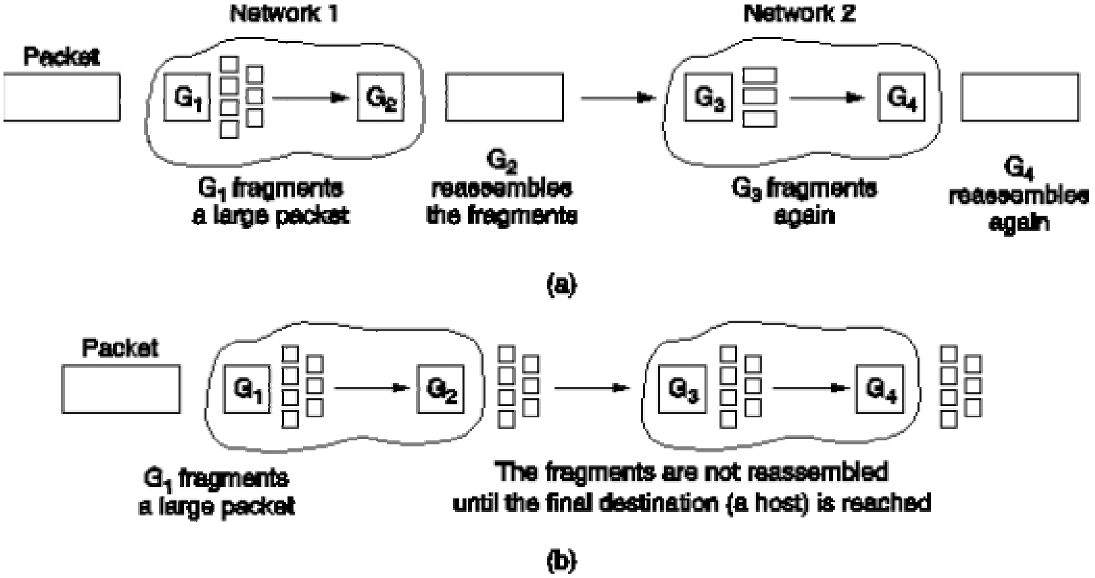

Network 1
Network 2
Packet
G_1
G_2
G_3
G_4
G_1 fragments a large packet
G_2 reassembles the fragments
G_3 fragments again
G_4 reassembles again
(a)
Packet
G_1
G_2
G_3
G_4
G_1 fragments a large packet
The fragments are not reassembled until the final destination (a host) is reached
(b)

Fig. 5-41. (a) Transparent fragmentation. (b) Nontransparent fragmentation.

When a packet is fragmented, the fragments must be numbered in such a way that the original data stream can be reconstructed.

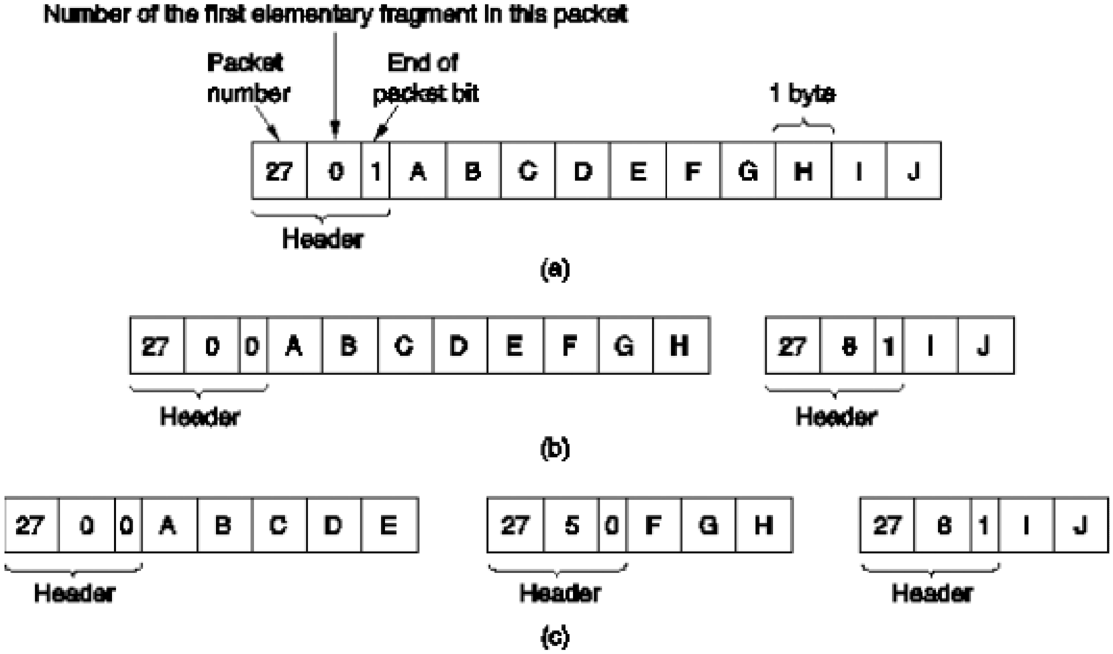


Fig. 5-42. Fragmentation when the elementary data size is 1 byte.
(a) Original packet, containing 10 data bytes.
(b) Fragments after passing through a network with maximum packet size of 8 bytes.
(c) Fragments after passing through a size 5 gateway.

3.1.7. Firewalls

The ability to connect any computer to any computer is a mixed blessing. For individuals at home, wandering around the Internet is lots of fun. For corporate security managers, it is a nightmare.

Mechanisms are needed to protect systems as much as possible against the unauthorized access. Firewalls are able to accomplish this goal.

Firewall is an electronic drawbridge, all traffic to or from a company's LAN is forced to go through (Fig. 5-43). The firewall in this configuration has two components: two routers that do packet filtering and an application gateway. Simpler configurations also exists, but the advantage of this design is that every packet must transit two filters and an application gateway to go in or out.

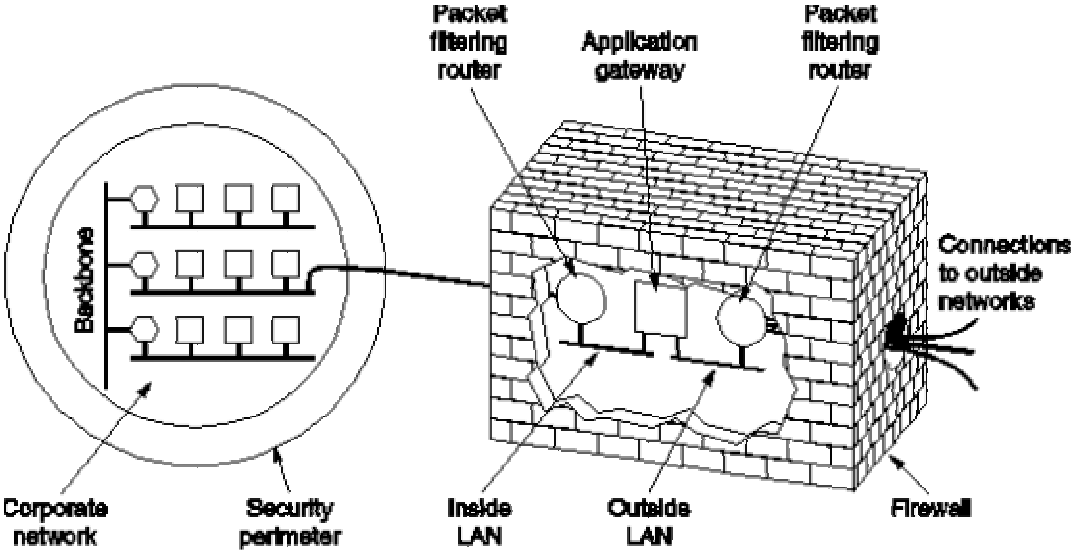


Fig. 5-43. A firewall consisting of two packet filters and an application gateway.

Each packet filter is a standard router equipped with some extra functionality allowing every incoming or outgoing packet to be inspected. Packets meeting some criterion are forwarded normally. Those that fail the test are dropped.

In Fig. 5-43, the packet filter on the inside LAN could check outgoing packets and the one on the outside LAN checks incoming packets. The point of putting the two packet filters on different LANs is to ensure that no packet gets in or out without having pass through the application gateway.

Packet filters are typically driven by tables configured by the system administrator. These tables list sources and destinations that are acceptable and blocked. In the common case of a UNIX setting, a source or destination consists of an IP address and a port. Ports indicate which service is desired. For example, port 23 is for Telnet, port 79 is for Finger. A company could block incoming packets for all IP addresses combined with one of these ports. In this way, no one outside the company could log in via Telnet, or look up people using the Finger daemon.

The second half of the firewall mechanism is the application gateway operating at the application level. A mail gateway, for example, can be set up to examine each message going in or coming out. For each one it makes a decision to transmit or discard it based on header fields, message size, or even content (a sensibility to some words can be set up).

Note: if some of the machines uses a wireless connection with the outside world, no firewall can ensure security of the network.

3.2. The Network Layer In The Internet

At the network layer, the Internet can be viewed as a collection of subnetworks or *Autonomous Systems* (ASes) that are connected together. There is no real structure, but several major backbones exist. These are constructed from high-bandwidth lines and fast routers. Attached to the backbones are regional (midlevel) networks, and attached to these regional networks are the LANs at many universities, companies, and Internet service providers (Fig. 5-44).

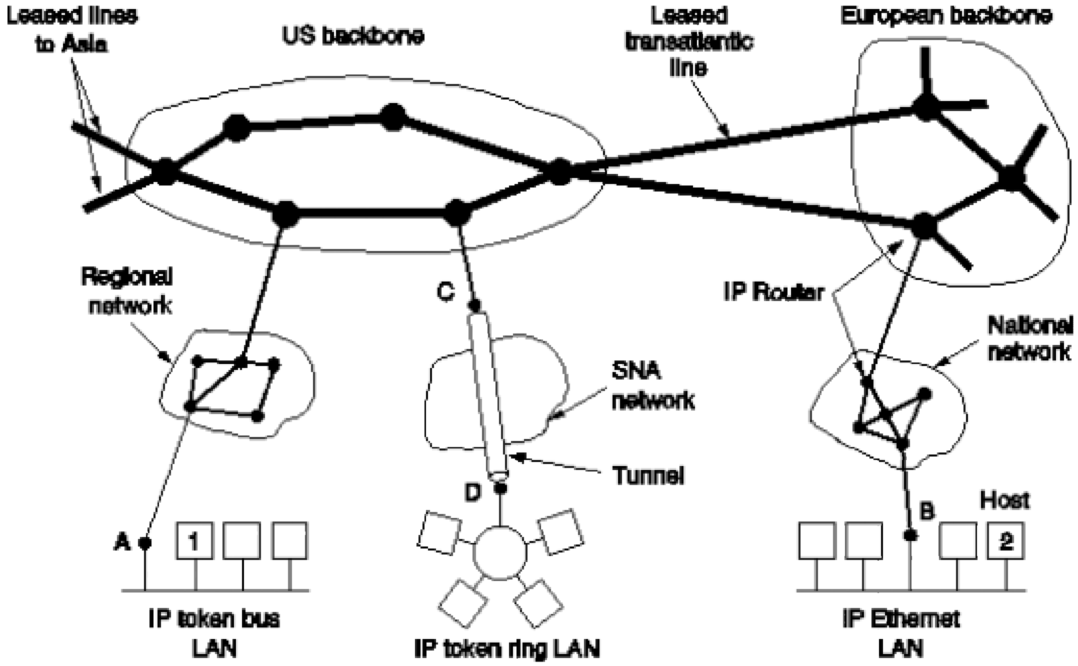


Fig. 5-44. The Internet is an interconnected collection of many networks.

The glue that holds the Internet together is the network layer protocol, *IP* (Internet Protocol).

Communication in the Internet works as follows: The transport layer takes data streams and break them up into datagrams usually around 1500 bytes long (in theory, they can be up to 64 Kbytes). Each datagram is transmitted through the Internet, possibly being fragmented into smaller units. When all pieces finally get to the destination machine, they are reassembled by the network layer into the original datagram. This datagram is then handed to the transport layer, which inserts it into the receiving process' input stream.

3.2.1. The IP Protocol

An IP datagram consists of a header part and a text part. The header has a 20 byte fixed part and a variable length optional part (Fig. 5-45). It is transmitted from left to right, with the high-order field of the Version field going first (big endian order).

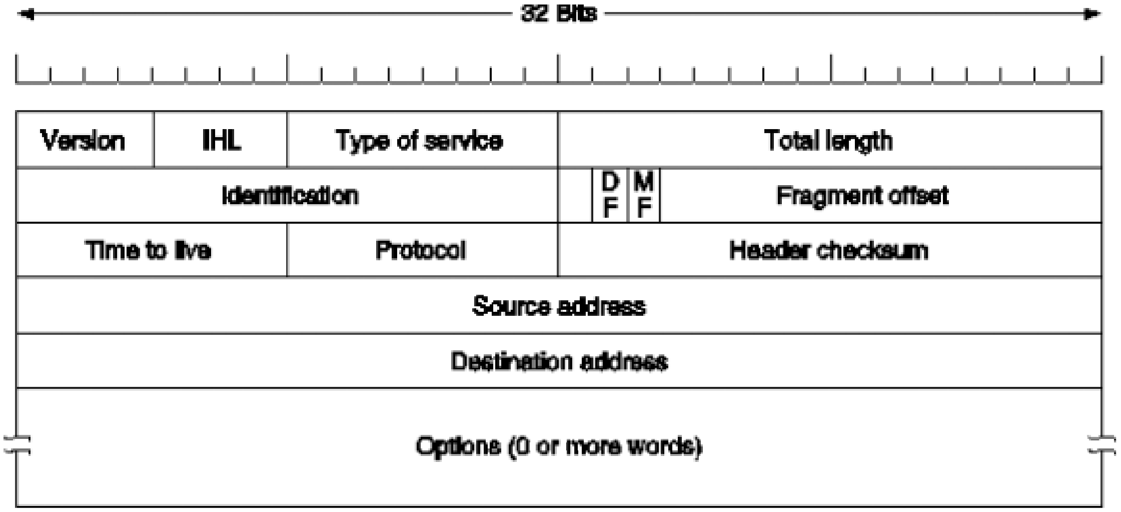
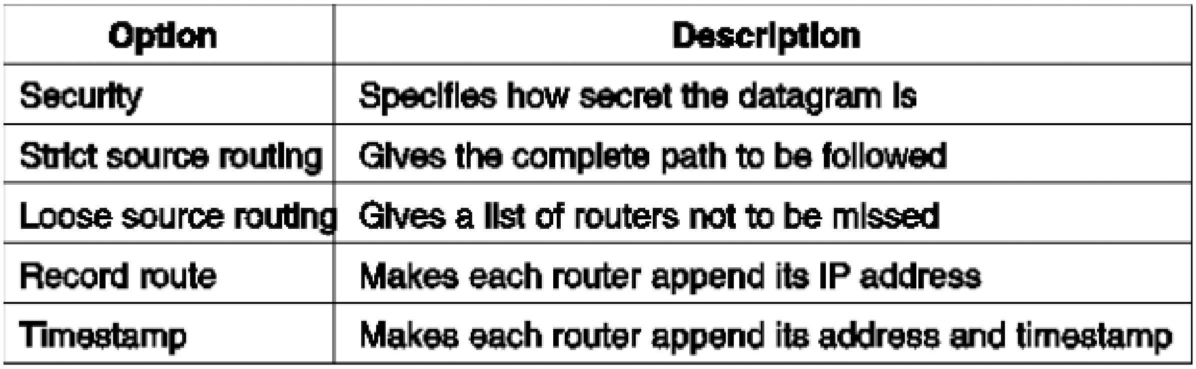


Fig. 5-45. The IP (Internet Protocol) header.

Meanings of single fields in the IP header:

- Version - version of the protocol the datagram belongs to.
- IHL - length of the header in 32-bit words. The minimum value is 5, the maximum value is 15.
- Type of service - type of service the host wants from a subnet. In practice, current routers ignore this field.
- Total length - the total length of the datagram (65535 bytes maximum).
- Identification - allows the destination host to determine which datagram a newly arrived fragment belongs to. All the fragments of a datagram contain the same Identification value.
- DF - stands for Don't fragment. It is an order to the router not to fragment the datagram because the destination is incapable of putting the pieces back together again (all machines are required to accept fragments of 576 bytes or less).
- MF - stands for More Fragments. All fragments except the last one have this bit set.
- Fragment offset - tells where in the current datagram this fragment belongs. All fragments except the last one in a datagram must be a multiple of 8 bytes, the elementary fragment unit.
- Time to live - a counter used to limit packet lifetimes. In practice, it just counts hops. It must be decremented on each hop. When it hits zero, the packet is discarded and a warning packet is sent back to the source host.
- Protocol - the identification of the protocol the datagram belongs to.
- Header checksum - verifies the header only. It is recomputed at each hop.
- Source address, destination address - indicate the network number and host number.
- Options - provides an escape to allow subsequent versions of the protocol to include information not present in the original design, to permit experiments to try out new ideas. This field is variable length. Each option begins with a 1-byte code identifying the option. Currently five options are defined, but not all routers support them. E.g., Record route option tells the routers along the path to append their IP address to the option field. This allows system managers to track down bugs in the routing algorithms.



Option	Description
Security	Specifies how secret the datagram is
Strict source routing	Gives the complete path to be followed
Loose source routing	Gives a list of routers not to be missed
Record route	Makes each router append its IP address
Timestamp	Makes each router append its address and timestamp

Fig. 5-46. IP options.

3.2.2. IP Addresses

Every host and router on the Internet has an *IP address*, which encodes its network number (prefix part of the address) and host number (suffix part of the address). The combination is unique.

All IP addresses are 32 bits long. The format of IP address is in Fig. 5-47. Those machines connected to multiple networks have a different IP address on each network.

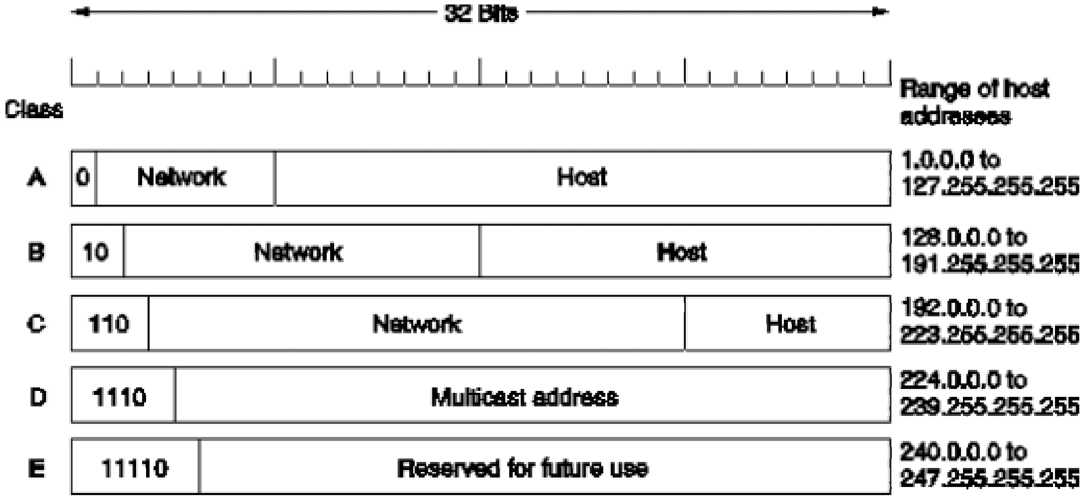

32 Bits
Class
Range of host addresses
A
0
Network
Host
1.0.0.0 to 127.255.255.255
B
10
Network
Host
128.0.0.0 to 191.255.255.255
C
110
Network
Host
192.0.0.0 to 223.255.255.255
D
1110
Multicast address
224.0.0.0 to 239.255.255.255
E
11110
Reserved for future use
240.0.0.0 to 247.255.255.255

Fig. 5-47. IP address formats.

IP addresses can be divided into 5 classes. There are 3 primary classes, A, B, and C, used for host addresses. Class D is used for multicasting which allows delivery to a set of computers. Class E is reserved for future use. The first four bits of an address determine the class to which the address belongs.

The *class A* format allows for up to 126 networks with 16 million hosts each.

The *class B* format allows for up to 16382 networks with 64 K hosts each.

The *class C* format allows for up to 2 million networks with 254 hosts each.

Network numbers on the top level are assigned by the NIC (Network Information Center). For single organizations, the network number are assigned by Internet service providers.

Network addresses are usually written in *dotted decimal notation*. In this format, each of the 4 bytes is written in decimal, from 0 to 255.

Some addresses have special meaning (Fig. 5-48).

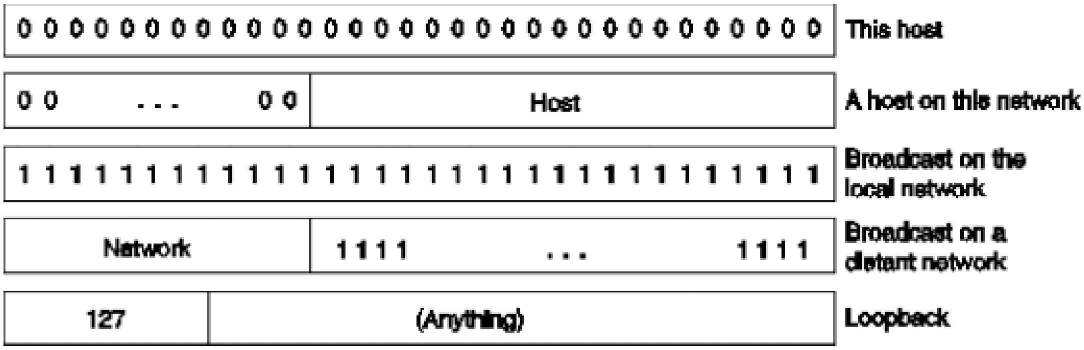


Fig. 5-48. Special IP addresses.

The IP address 0.0.0.0 is used by hosts when they are being booted but is not used afterwards.

All addresses of the form 127.xx.yy.zz are reserved for loopback testing. Packets sent to that address are not put onto the wire; they are processed locally and treated as incoming packets.

3.2.3. Subnets

All the hosts in a network must have the same network numbers. When the number of computers in an organization get bigger or the number of different LANs get bigger, this requirement can cause problems.

The solution to these problems is to allow a network to be split into several parts for internal use but still acts like a single network to the outside world. In the Internet literature, these parts are called *subnets* (different from subnets as a collections of routers).

The division is done in fact by splitting the host part of the address (e.g., 16 bits in case of B address) into a subnet number (e.g., 6 bits) and a host number (10 bits in our example). This split allows 62 LANs (0 and 255 are reserved), each with up to 1022 hosts (Fig. 5-49).

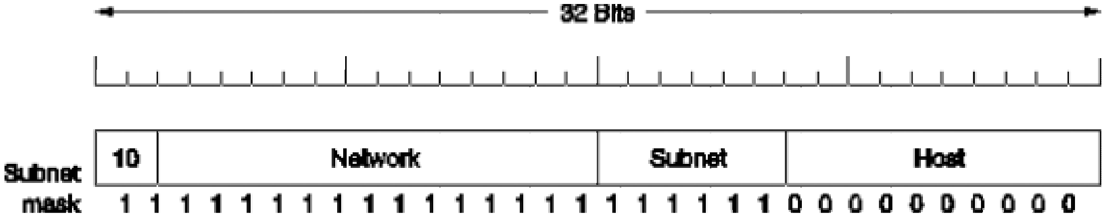


Fig. 5-49. One of the ways to subnet a class B network.

To see how subnets work, it is necessary to explain how IP packets are processed at a router.

Each router has a table listing some number of (network, 0) IP addresses and some number of (this-network, host) IP addresses. The first kind tells how to get to distant networks. The second kind tells how to get to local hosts. Associated with each table is the network interface to use to reach the destination, and certain other information.

When an IP packet arrives, its destination address is looked up in the routing table. If the packet is for a distant network, it is forwarded to the next router on the interface given in the table. If it is a local host (e.g., on the router LAN), it is sent directly to the destination. If the network in the destination address is not present in the router's table, the packet is forwarded to a default router with more extensive tables. So each router has to keep track of other networks and local hosts, not (network,host) pairs.

When subnetting is introduced, the routing tables are changed, adding entries of the form (this-network, subnet, 0) and (this-network, this-subnet, host). Thus a router on subnet k knows how to get to all other subnets and also how to get to all the hosts on subnet k. In fact, all that needs to be changed is to have each router do a Boolean AND with the network's subnet mask (Fig. 5-49) to get rid of the host number and look up the resulting address in its tables (after determining which network class it is). Subnetting reduces router table space by creating a three-level hierarchy.

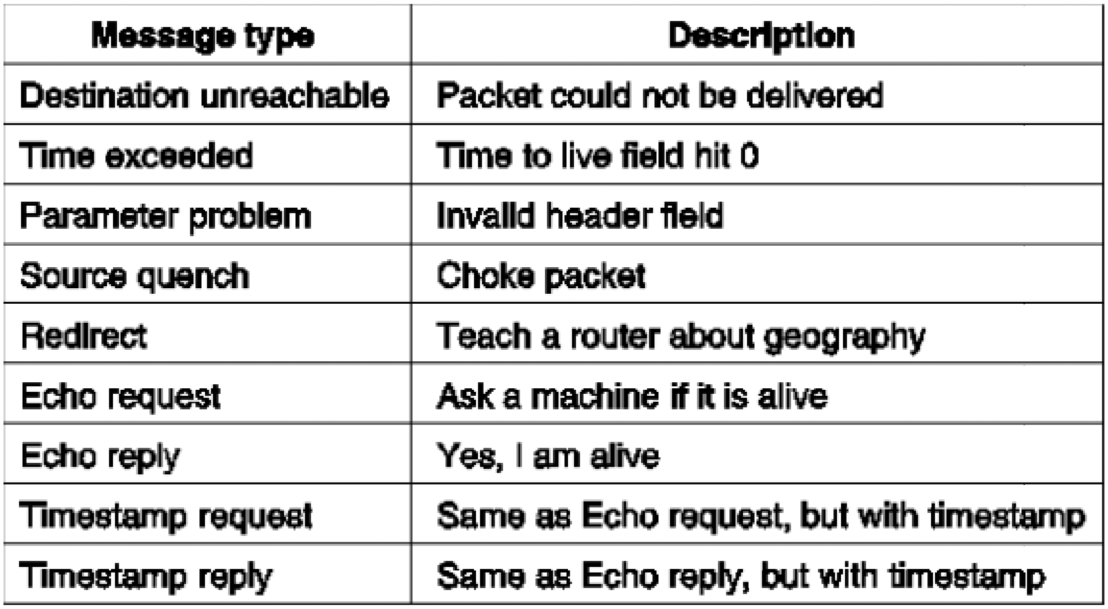
3.2.4. Internet control protocols

In addition to IP, which is used for data transfer, the Internet has several control protocols used in the network layer, including ICMP and ARP.

3.2.5. The Internet Control Message Protocol

The operation of the Internet is monitored by the routers. When something unexpected occurs, the event is reported by the ICMP (Internet Control Message Protocol).

Each ICMP message is encapsulated in an IP packet. The most important messages are in Fig. 5-50.



Message type	Description
Destination unreachable	Packet could not be delivered
Time exceeded	Time to live field hit 0
Parameter problem	Invalid header field
Source quench	Choke packet
Redirect	Teach a router about geography
Echo request	Ask a machine if it is alive
Echo reply	Yes, I am alive
Timestamp request	Same as Echo request, but with timestamp
Timestamp reply	Same as Echo reply, but with timestamp

Fig. 5-50. The principal ICMP message types.

The more detailed meaning of single messages are as follows:

- DESTINATION UNREACHABLE - a router cannot locate the destination, or a packet with the DF bit cannot be delivered because a "small-packet" network stands in the way.
- TIME EXCEEDED - a packet is dropped due to its router reaching zero.
- PARAMETER PROBLEM - an illegal value has been detected in a header file.
- SOURCE QUENCH - used formerly to throttle hosts that were sending too many packets. Today congestions are solved by another means.
- REDIRECT - a router notices that a packet seems to be routed wrong.

- ECHO REQUEST, ECHO REPLY - messages used to see if a given destination is reachable and alive. Upon receiving an ECHO REQUEST message, the destination is expected to send an ECHO REPLY message back.
- TIMESTAMP REQUEST, TIMESTAMP REPLY - similar to echo messages, except that the arrival time of the message and the departure time of the reply are recorded in the reply. This facility is used to measure network performance.

The ICMP is defined in RFC 792.

3.2.6. The Address Resolution Protocol

The ARP (Address Resolution Protocol), defined in RFC 826, solves the following problem: A computer on a LAN has to send an IP packet with the destination IP address A to a computer on the same LAN (the fact that the computer with the address A is on the same LAN is known from the address A), but it does not know the LAN address of the computer necessary to send a packet directly. So it, using ARP, broadcast packet on the LAN asking: Who owns IP address A? The broadcast will arrive at every machine on the LAN and each one will check its IP address. The host with the IP address A will respond announcing its LAN address. An example is in Fig. 5-51.

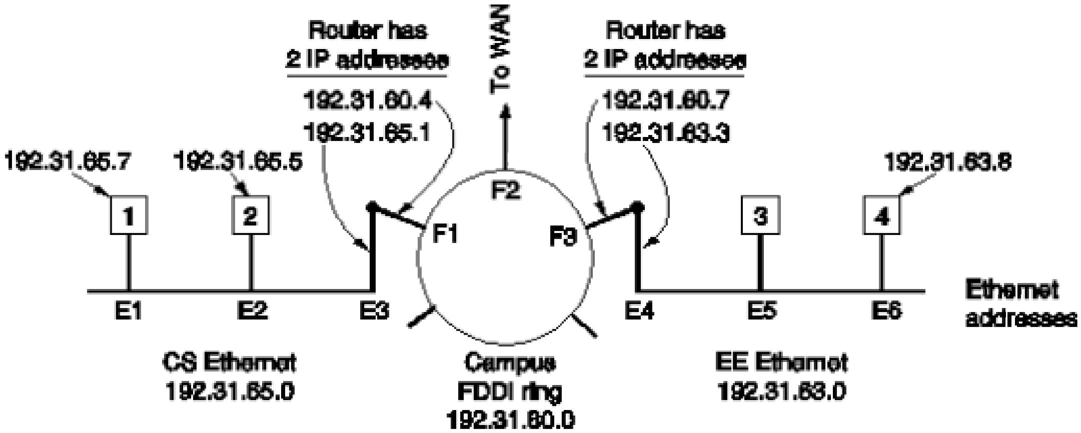


Fig. 5-51. Three interconnected class C networks: two Ethernets and an FDDI ring.

3.2.7. The Interior Gateway Routing Protocol: OSPF

Today, the Internet is made up of large number of autonomous systems (AS). Each AS is operated by a different organization and can use its own routing algorithm inside.

A routing algorithm within an AS is called an *interior gateway protocol*. An algorithm for routing between ASes is called *exterior gateway protocol*.

The original Internet interior gateway protocol was a distance vector protocol (*Routing Information Protocol* - RIP) based on Bellman-Ford algorithm. Gateways broadcast messages containing information from their routing databases. Each message consists of pairs, where each pair contains an IP network address and an integer distance to that network in hop count metrics. Other gateways listen to the broadcasted messages and update their tables according to the vector-distance algorithm. This protocol worked well in small systems, but less well as ASes got larger.

In 1979, the RIP was replaced by a link state protocol based on computing the shortest path to single routers. In 1990, a successor of this protocol, *OSPF* (Open Shortest Path First), became a standard (RFC 1247).

Requirements taken into account when OSPF was designed:

- the algorithm has to be published in open literature,
- it has to support the variety of distance metrics,
- it has to be based on dynamic algorithm,
- it has to support routing based on types of services,
- it has to do load balancing, splitting the load over multiple lines,
- support for hierarchical systems is needed, no router could be expected to know the entire topology,
- some extent of security is required,
- provision is needed for dealing with routers connected to the Internet via a tunnel.

OSPF supports three kinds of connections and networks:

1. Point-to-point lines between exactly two routers.
2. Multiaccess networks with broadcasting (most LANs).
3. Multiaccess networks without broadcasting (most packet switched networks).

A multi-access network is one that can have multiple routers on it, each of which can directly communicate with all others. All LANs and WANs have this property (Fig. 5-52(a)). Hosts do not generally play a role in OSPF.

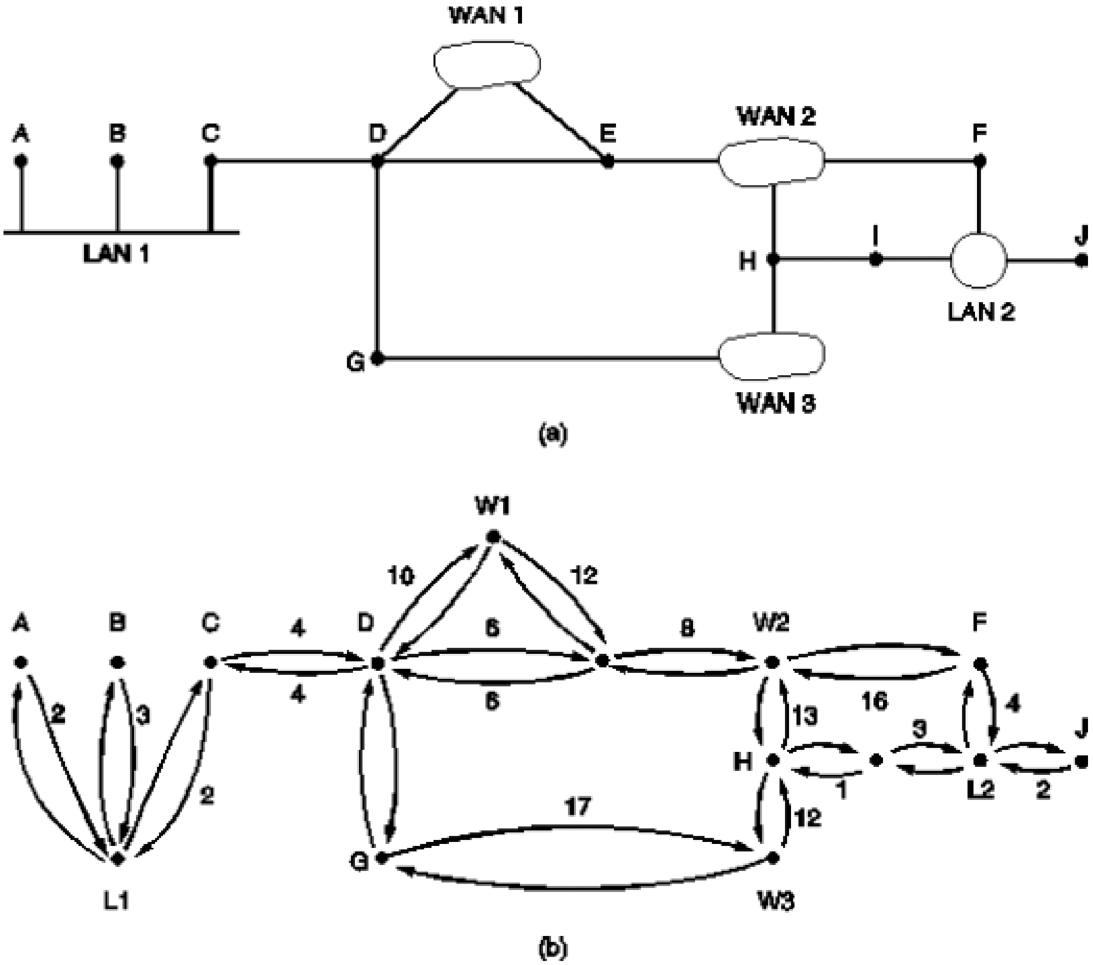


Fig. 5-52. (a) An autonomous system. (b) A graph representation of (a).

OSPF works by abstracting the collection of actual networks, routers, and lines into a directed graph in which each arc is assigned a cost. It then computes the shortest path based on the weight on the arcs. A serial connection between two routers is represented by a pair of arcs one in each direction. Their weights may be different.

A multi-access network is represented by a node for the network itself plus a node for each router. The arc from the network node to the routers have weight 0 and are omitted from the graph. Fig. 5-52(b) shows the graph representation of the network of Fig. 5-52(a).

OSPF represents the actual network as a graph and then compute the shortest path from every router to every other router.

OSPF allows to divide ASes into numbered areas. Area is a network or a set of contiguous networks. Areas do not overlap but need not be exhaustive, i.e., some routers may belong to no area. An area is a generalization of a subnet. Outside an area, its topology and details are not visible.

Every AS has a backbone area called area 0. All areas are connected to the backbone. Each router that is connected to two or more areas is part of the backbone.

Within an area, each router has the same link state database and runs the same shortest path algorithm.

OSPF handles different types of service routing in a different way. For each service it maintains the corresponding graph and makes special computations.

During the normal operation, three kinds of routes may be needed:

- intra-area - here the source router already knows the shortest path to the destination router,
- inter-area,
- inter AS.

Inter-area routing proceeds in three steps:

1. go from the source to the backbone,
2. go across the backbone to the destination area,
3. go to the destination.

Packets are routed from source to destination "as is", they are not encapsulated or tunneled unless going to an area whose only connection to the backbone is a tunnel.

Fig. 5-53 shows part of Internet with ASes and areas.

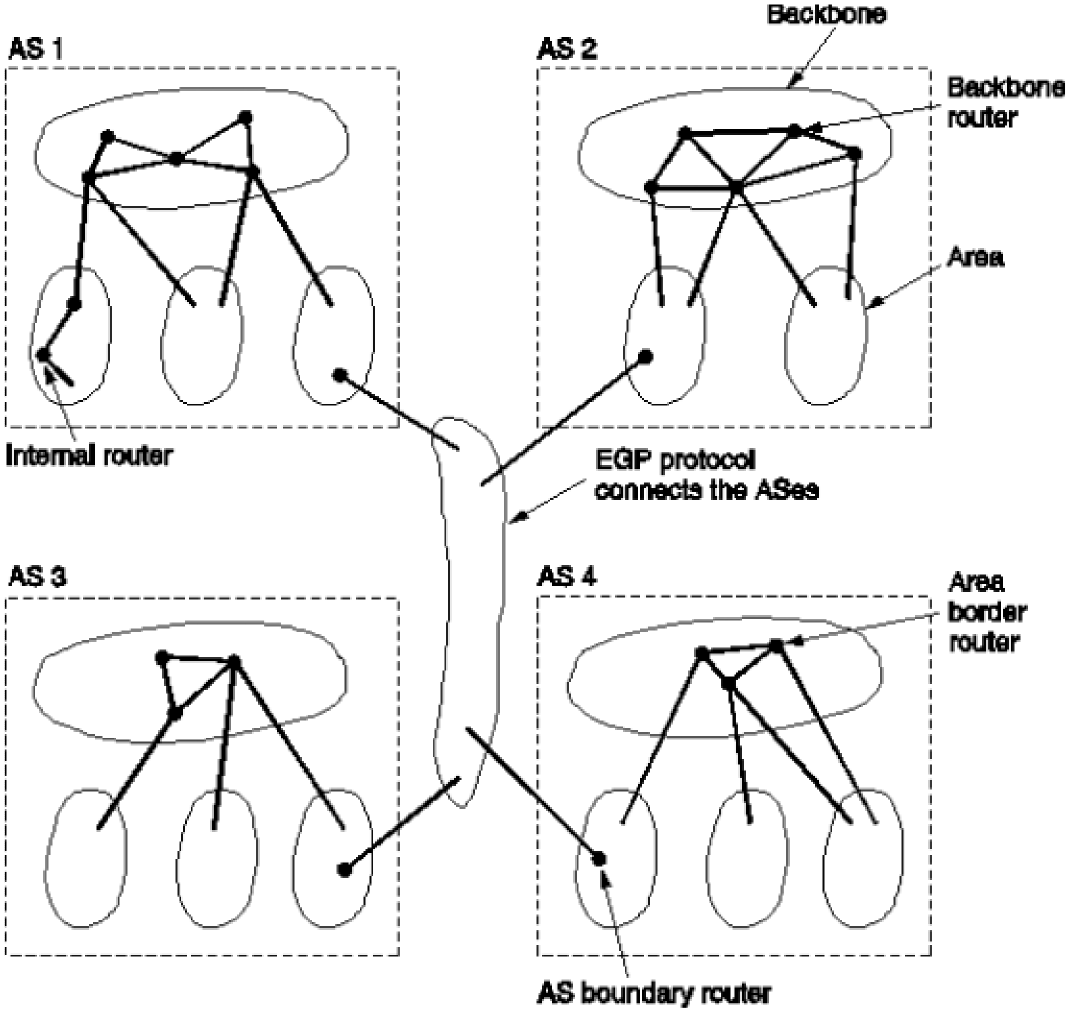


Fig. 5-53. The relation between ASes, backbones, and areas in OSPF.

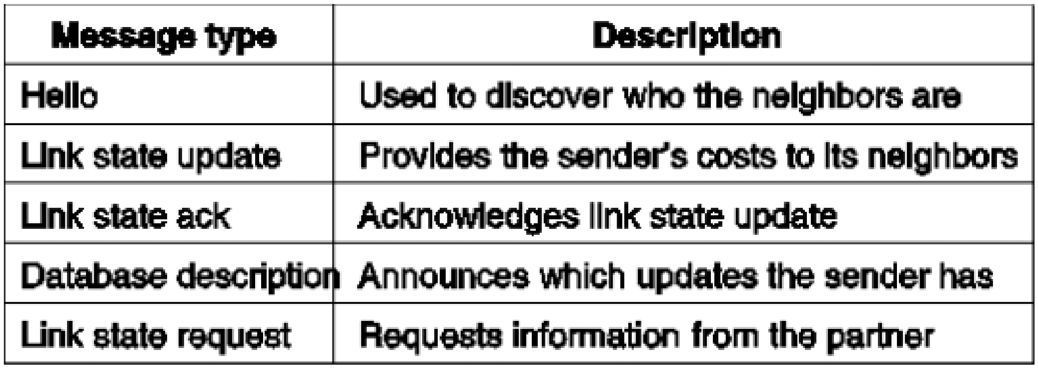
OSPF distinguishes four classes of routers:

- internal routers are wholly within one area,
- area border routers connect two or more areas,
- backbone routers are on the backbone,
- AS boundary routers talk to routers in other ASes.

These classes are allowed to overlap.

When a router boots, it sends HELLO messages on all of its point-to-point lines and multicasts them on LANs to the group consisting of all other routers. From the responses, each router learns who its neighbors are.

OSPF works by exchanging information between adjacent routers which is not the same as between neighboring routers. On each LAN, one router is selected as designated router. It is said to be adjacent to all other router and exchanges information with them. A backup designated router is always kept up-to-date to easy transition should the primary designated router crash.



Message type	Description
Hello	Used to discover who the neighbors are
Link state update	Provides the sender's costs to its neighbors
Link state ack	Acknowledges link state update
Database description	Announces which updates the sender has
Link state request	Requests information from the partner

Fig. 5-54. The five types of OSPF messages.

During normal operation, each router periodically floods LINK STATE UPDATE messages to each of its adjacent routers. This message gives its state and provides the costs used in topological database. These messages are acknowledged to make them reliable. Each message has a sequence number, so a router can see whether an incoming LINK STATE UPDATE is older or newer that what it currently has. Routers also send these messages when a line goes up or down or its cost changes.

DATABASE DESCRIPTION messages give the sequence numbers of all the link state entries currently held by the sender. By comparing its own values with those of the sender, the receiver can determine who has the most recent values.

Either partner can request link state information from the other one using LINK STATE REQUEST messages. So each pair of adjacent routers checks to see who has the most recent data, and new information is spread throughout the area this way.

So, the complete picture is the following: Using flooding, each router informs all the other routers in its area of its neighbors and costs. This information allows each router to construct the graph for its area(s) and compute the shortest path. The backbone area does this too.

3.2.8. The Exterior Gateway Routing Protocol: BGP

Exterior gateway protocol is used between ASes and unlike interior gateway protocol it has to worry about politics. For example, a corporate AS might be unwilling to carry transit packets originating in a foreign AS and ending in a different foreign AS. Exterior gateway protocols in general, and BGP in particular, have been designed to allow many kinds of routing policies to be enforced in the inter AS traffic.

Typical policies involve political, security, or economic considerations. Some examples:

1. No transit traffic through certain ASes.
2. Never put Iraq on a route starting at the Pentagon.
3. Do not use the United States to get from British Columbia to Ontario.
4. Only transit Albania if there is no alternative to the destination.
5. Traffic starting or ending at IBM should not transit Microsoft.

Policies are manually configured into each BGP router. They are not part of the protocol itself.

From the point of view of a BGP router, the world consists of other BGP routers and the lines connecting them. Two BGP routers are considered connected if they share a common network.

Given BGP's special interest in transit traffic, networks are grouped into one of three categories:

- *Stub networks* - they have only one connection to the BGP graph. Therefore, these cannot be used for transit traffic.
- *Multiconnected networks* - they could be used for transit traffic, except that they refuse.
- *Transit networks* - such as backbones, they are willing to handle third-party packets, possibly with some restrictions.

Pairs of BGP routers communicate with each other by establishing TCP connections. Operating this way provides reliable communication and hides all the details of the network being passed through.

BGP is fundamentally a distance vector protocol, but quite different from most other such as RIP. Instead of maintaining just the cost to each destination, each BGP router keeps track of the exact path used. Similarly, instead of periodically giving each neighbour its estimated cost to each possible destination, each BGP router tells its neighbours the exact path it is using (Fig. 5-55). Every BGP router contains a module that examines routes to a given destination and scores them, returning a number for the "distance" to that destination for each route. Any route violating a policy constraint automatically gets a score of infinity. The router then adopts the route with the shortest distance. The scoring function is not part of the BGP protocol and can be any function the system manager wants.

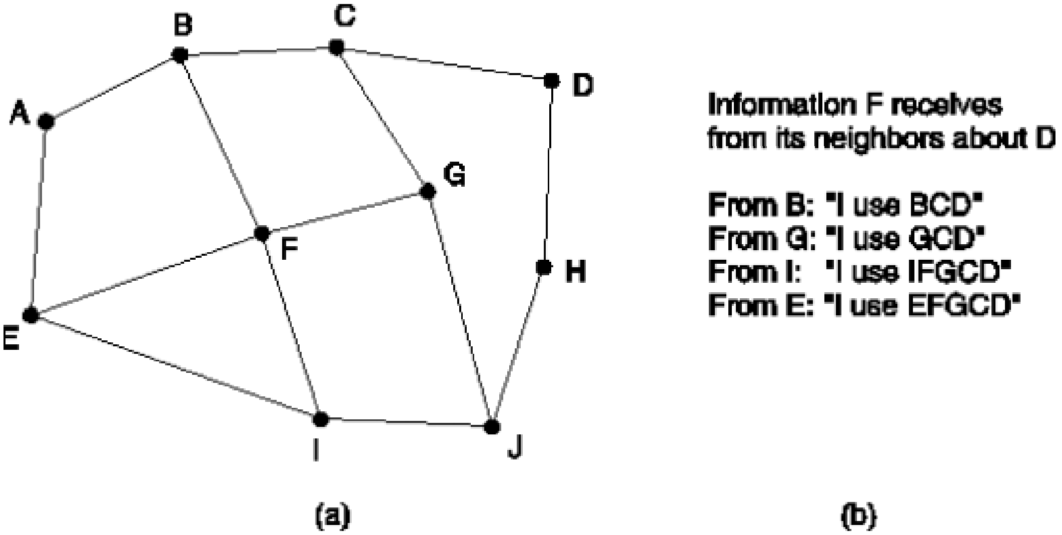


Fig. 5-55. (a) A set of BGP routers. (b) Information sent to F.

The current definition of BGP is in RFC 1654. Additional useful information can be found in RFC 1268.

3.2.9. CIDR - Classless InterDomain Routing

IP has worked extremely well, as demonstrated by the exponential growth of the Internet. Unfortunately, IP is rapidly becoming a victim of its own popularity: it is running out of addresses.

In principle, over 2 billion addresses exist, but the practice of organizing the address space by classes wasted millions of them. The biggest *problems* are with class B network. For most organizations, a class A network, with 16 million addresses is too bog, and a class C network, with 256 addresses is too small. So they asked for class B network having 65536 addresses. In reality, a class B address is far too large for most organizations. In retrospect, it might have been better to have had class C networks use 10 bits instead of eight for the host number.

There is also another problem: the size of routing tables and the increasing complexity of routing algorithms. There are many solutions proposed but usually they solve one problem and create a new one.

One solution that is now being implemented and which will give Internet a bit of extra breathing room is CIDR - Classless InterDomain Routing. The basic idea behind CIDR (described in RFC 1519), is to allocate the remaining class C networks, of which there are almost 2 million, in variable-size blocks. If a site needs, say, 2000 addresses, it is given a block of 2048 addresses (eight continuous class C networks), and not a full class B address.

The allocation rules for the class C addresses were also changed (RFC 1519). The world was partitioned into four zones, and each one given a portion of the class C address space. The allocation was as follows:

Addresses 194.0.0.0 to 195.255.255.255 are for Europe.

Addresses 198.0.0.0 to 199.255.255.255 are for North America.

Addresses 200.0.0.0 to 201.255.255.255 are for Central and South America.

Addresses 202.0.0.0 to 203.255.255.255 are for Asia and the Pacific.

In this way, each region was given about 32 million addresses to allocate, with another 320 million class C addresses from 204.0.0.0 through 223.255.255.255 held in reserve for the future. The advantages of this allocation is that now any router outside Europe that gets a packet addressed to 194.xx.yy.zz or 195.xx.yy.zz can just send it to the standard European gateway having 32 million addresses now compressed into one routing table entry.

Of course, once a 194.xx.yy.zz packet gets to Europe, more detailed routing tables are needed. One possibility is to have 131072 entries for networks 194.0.0.xx through 195.255.255.xx, but this is precisely this routing table explosion we are trying to avoid. Instead, each routing table entry is extended by giving it a 32-bit mask. When a packet comes in, the routing table is scanned entry by entry, masking the destination address and comparing it to the table entry looking for a match.

Example: Let Cambridge University needs 2048 addresses and it is assigned the addresses 194.24.0.0 through 194.24.7.255, along with a mask 255.255.248.0. Next, Oxford University asks for 4096 addresses. Since a block of 4096 addresses must lie on a 4096-byte boundary, they cannot be given addresses starting at 194.24.8.0. Instead they get 194.24.16.0 through 194.24.31.355 along with mask 255.255.240.0.Now the University of Edinburg asks for 1024 addresses and is assigned addresses 194.24.8.0 through 194.24.11.255 and mask 255.255.252.0.

The routing tables all over Europe are now updated with the following entries:

Address	Mask
11000010 00011000 00000000 00000000	11111111 11111111 11111000 00000000
11000010 00011000 00010000 00000000	11111111 11111111 11110000 00000000
11000010 00011000 00001000 00000000	11111111 11111111 11111100 00000000

When a packet addressed 194.24.17.4, which in binary is:

11000010 00011000 00010001 00000100

comes in, first, it is Boolean ANDed with the Cambridge mask to get

11000010 00011000 00010000 00000000

which does not match the Cambridge base address, so the original address is next ANDed with the Oxford mask to get

11000010 00011000 00010000 00000000

This value does match the Oxford base address, so packet is sent to the Oxford router. It is possible for two entries to match, in which case the one whose mask has the most 1 bits wins. The same idea can be applied to all addresses, not just the new class addresses, so with CIDR, the old class A, B, and C networks are no longer used for routing. This is why CIDR is called classless routing.

3.2.10. User Datagram Protocol

At the IP layer, a destination address identifies a host computer, no further distinction is made regarding which user or which application program will receive the datagram. User Datagram Protocol (UDP) provides one of possible mechanisms allowing multiple application programs executing on a given computer to send and receive datagrams independently.

3.2.11. Identifying The Ultimate Destination

The operating systems in most computers support multiprogramming, which means they permit multiple application programs to execute simultaneously. We refer to each executing program as a process. It may seem natural to say that a process is the ultimate destination for a message. But there are several problems with such an approach.

Instead of thinking of a process as the ultimate destination, we will imagine that each machine contains a set of abstract destination points called protocol ports. Each protocol port is identified by a positive integer. The local operating system provides an interface mechanism that processes use to specify a port or access it.

Most operating systems provide synchronous access to ports. From a particular process point of view it means that the computation stops during a port access operation. For example, if a process attempts to extract data from a port before any data arrives, the operating system stops the process until data arrives. Once the data arrives, the operating system passes the data to the process and restarts it.

In general, ports are buffered, so data that arrives before a process is ready to accept it will not be lost.

To communicate with a foreign port, a sender needs to know both the IP address of the destination machine and the protocol port number of the destination within that machine.

3.2.12. The User Datagram Protocol

UDP belongs to TCP/IP protocol suite. It provides the primary mechanism that application programs use to send datagrams to other application programs. UDP provides protocol ports used to distinguish among multiple processes executing on a single machine. Each UDP message contains both a destination port number and a source port number, making it possible for the UDP software on the destination to deliver the message to the correct recipient to send a reply.

UDP uses the underlying Internet Protocol to transport a message from one machine to another, and provides the same unreliable, connectionless datagram delivery service as IP. It just adds the ability to distinguish among multiple destinations within a given host computer.

An application program that uses UDP accepts full responsibility for handling the problem of reliability.

3.2.13. Format of UDP Messages

Each UDP message is called a *user datagram*. It consists of UDP header and UDP data area.

The header consists of four 16-bit fields that specify the port from which the message was sent, the port to which the message is destined, the message length (in octets), and a UDP checksum. The source port is optional, if not used, it should be zero. The checksum is also optional, if not used, it should be zero.

3.2.14. UDP Encapsulation and Protocol Layering

UDP is a protocol of the transport layer. Conceptually, application programs access UDP, which uses IP to send and receive datagrams.

Layering UDP above IP means that a complete UDP message, including the UDP header and data, is encapsulated in an IP datagram as it travels across an internet. The IP layer is responsible only for transferring data between a pair of hosts on an internet, while UDP layer is responsible only for differentiating among multiple sources or destinations within one host.

3.2.15. Reserved and Available UDP Port Numbers

There are two fundamental approaches to port assignment.

The first approach uses central authority. Everyone agrees to allow a central authority to assign port numbers as needed and to publish the list of assignments. Then all software is built according to the list. This approach is sometimes called universal assignment and the port assignments specified by the authority are called well-known port assignments.

The second approach uses dynamic binding. Here ports are not globally known. Instead, whenever a program needs a port the network software assigns one. To learn about the current port assignment on another computer, it is necessary to send a request that asks a question like "How do I reach the file transfer service?". The target machine replies by giving the correct port number to use.

The TCP/IP designers adopted a hybrid approach that assigns a few port numbers a priori, but leaves most available for local sites or application programs. The assigned port numbers begin at low values and extend upward, leaving large integer values available for dynamic assignment.

3.2.16. The Internet Transport Protocol TCP

The Internet has two main protocols in the transport layer, a connection-oriented protocol TCP and a connectionless protocol UDP.

TCP (Transmission Control Protocol) was specifically designed to provide a reliable end-to-end byte stream over an unreliable network.

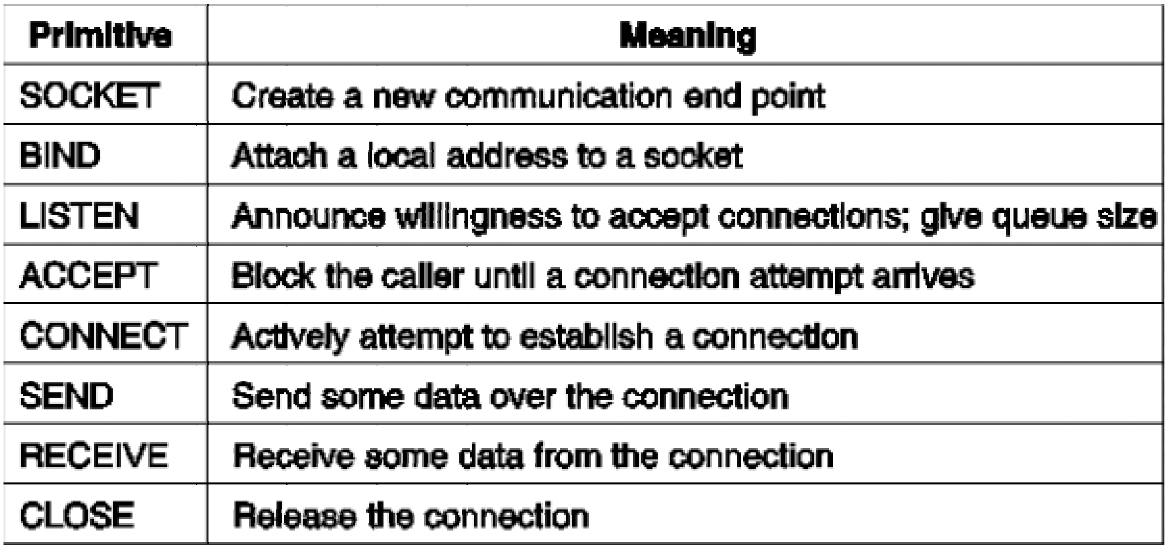
TCP was formally defined in RFC 793. As time went on, various errors were detected, and the requirements were changed. These clarifications are detailed in RFC 1122. Extensions are given in RFC 1323.

Each machine supporting TCP has a TCP transport entity, either a user process or part of the kernel that manages TCP streams and interfaces to the IP layer. A TCP entity accepts user data stream from local processes, break them up into pieces not exceeding 64K bytes (in practice, usually about 1500 bytes), and sends each piece as a separate IP datagram. When IP datagrams containing TCP data arrive at a machine, they are given to the TCP entity, which reconstructs the original byte streams.

The IP layer gives no guarantee that datagrams will be delivered properly, so it is up to TCP to time out and retransmit them as need to be. Datagrams that do arrive may well do so in the wrong order, it is also up to TCP to reassemble them into messages in the proper sequence. In short, TCP must furnish the reliability that most users want and that IP does not provide.

3.2.17. The TCP Service Model

TCP service is obtained by having both the sender and receiver create end points, called *sockets*. Each socket has a socket number (address) consisting of the IP address of the host and a 16-bit number local to that host, called a port. To obtain TCP service, a connection must be explicitly established between a socket on the sending machine and a socket on the receiving machine. The socket calls are listed in Fig. 6-6.



Primitive	Meaning
SOCKET	Create a new communication end point
BIND	Attach a local address to a socket
LISTEN	Announce willingness to accept connections; give queue size
ACCEPT	Block the caller until a connection attempt arrives
CONNECT	Actively attempt to establish a connection
SEND	Send some data over the connection
RECEIVE	Receive some data from the connection
CLOSE	Release the connection

Fig. 6-6. The socket primitives for TCP.

A socket may be used for multiple connections at the same time. In other words, more connections may terminate at the same socket. Connections are identified by the socket identifiers at both ends, that is, (socket1, socket2).

Port numbers below 256 are called well-known ports and are reserved for standard services. For example, any process wishing to establish a connection to a host to transfer a file using FTP can connect to the destination host's port 21 to contact its FTP daemon. To establish a remote login session using TELNET, port 23 is used. The list of well-known ports is given in RFC 1700.

All TCP connections are full-duplex and point-to-point. TCP does not support broadcasting and multicasting.

A TCP connection is a byte stream, not a message stream. Message boundaries are not preserved end-to-end. For example, if the sending process does four 512-bytes writes to a TCP stream, this data may be delivered to the receiving process as four 512-bytes chunks, two 1024-bytes chunks, or some other way. There is no way for the receiver to detect the units in which the data were written.

When an application passes data to TCP, TCP may send it immediately or buffer it (in order to collect a larger amount to send at once), at its discretion. If an application wants the data to be sent immediately (e.g. TELNET), it can use the PUSH flag, which tells TCP not to delay the transmission.

TCP also recognizes *urgent data*. When an interactive user hits CTRL-C key to break off a remote computation, the sending application puts the appropriate data to the TCP along with the URGENT flag. This causes TCP to transmit the data immediately. When the urgent data are received at the destination, the receiving application is interrupted and read the data stream to find the urgent data. The end of the urgent data is marked, so the application knows when it is over. The start of the urgent data is not marked. It is up to the application to figure that out.

3.2.18. The TCP Protocol

Every byte on a TCP connection has its own 32-bit sequence number. The sequence numbers are used both for acknowledgments and for the window mechanism, which use separate 32-bit header fields.

The sending and receiving TCP entities exchange data in the form of segments. A segment consists of a fixed 20-byte header (plus an optional part) followed by zero or more data bytes. The TCP software decides how big segments should be. It can accumulate data from several writes into one segment or split data from one write over multiple segments. Each segment including the TCP header must fit in the 65535 byte IP payload.

The basic protocol used by TCP entities is the sliding window protocol. When a sender transmits a segment, it also starts a timer. When the segment arrives at the destination, the receiving TCP entity sends back a segment bearing an acknowledgment number equal to the next sequence number it expects to receive. "Sliding window" means that the sender is allowed to send without acknowledgment all bytes from the stream laying in a predefined window that is sliding ahead as the acknowledgment are coming. If the sender's timer goes off before the acknowledgment is received, the sender transmits the segment again. This sounds simple but there are many problems that can occur and the TCP has to solve them.

3.2.19. The TCP Segment Header

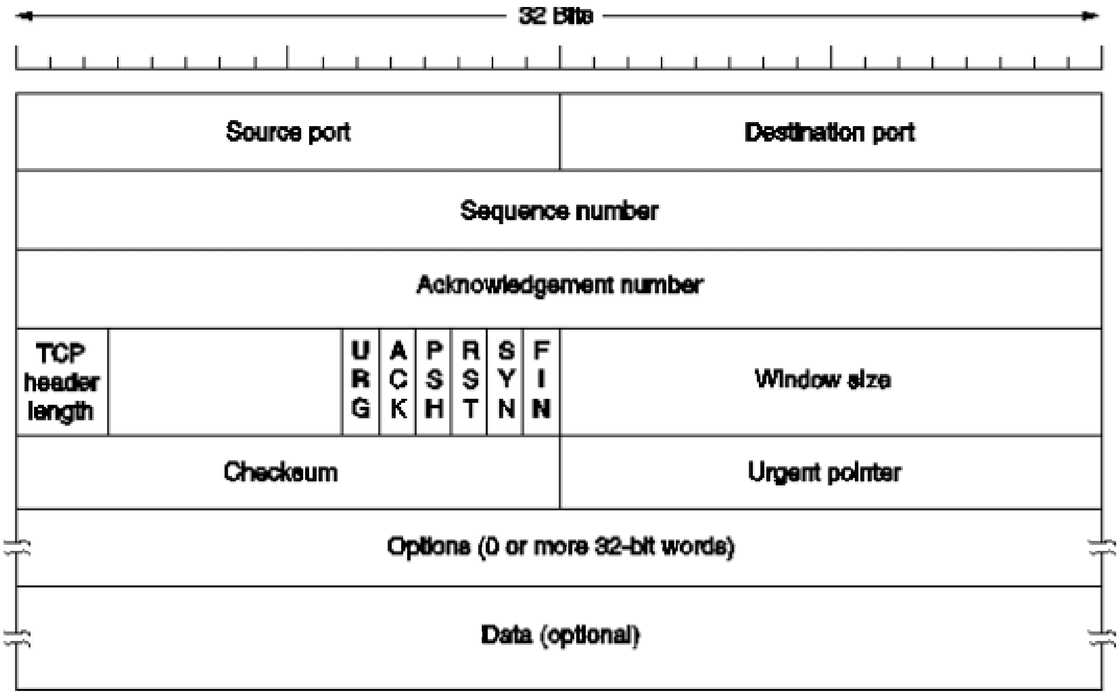


Fig. 6-24. The TCP header.

Fig. 6-24 shows the layout of a TCP segment. Its header contains the following fields:

- The Source port and Destination port - identify the local end points of the connection. Each host may decide for itself how to allocate its own ports starting at 256.
- The Sequence number and Acknowledgment number - perform their usual functions. The former indicates the byte sequence number of the first octet in this TCP data block and is incremented according to the number of octets transmitted in each TCP segment, the latter specifies the next byte expected, not the last byte correctly received.
- The TCP header length - the number of 32-bit words contained in the TCP header.
- 6-bits unused.
- Six 1-bit flags. URG is set to 1 if the Urgent pointer is in use. The Urgent pointer is used to indicate a byte offset from the current sequence number at which urgent data are to be found. The ACK bit is set to 1 to indicate that the Acknowledgment number is valid. If ACK is 0, the segment does not contain an acknowledgment. The PSH bit indicates PUSHed data. The receiver is hereby requested to deliver the data to the application upon arrival and not to buffer it. The RST bit is used to reset a connection that has become confused due to a host crash or some other reason. It is also used to reject an invalid segment or refuse an attempt to open a connection. The SYN bit is used to establish connections. It is used to denote CONNECTION REQUEST when ACK = 0 and CONNECTION ACCEPTED when ACK = 1. The FIN bit is used to release a connection. It specifies that the sender has no more data to transmit.
- The Window field tells how many bytes may be sent starting at the byte acknowledged. If equals 0, it says that the bytes up to and including Acknowledgment number - 1 have been received, but that the receiver would like no more data for the moment. Permission to send can be granted later by sending a segment with the same Acknowledgment number and nonzero Window field.
- Checksum checksums the header, the data, and the conceptual pseudoheader shown in Fig. 6-25.
- The Option field was designed to provide a way to add extra facilities not covered by the regular header. The most important option is the one that allows each host to specify the maximum TCP payload it is willing to accept. During connection setup, each side can announce its maximum and see its partner’s. The smaller of the two numbers wins. If a host does not use this option, it defaults to a 536-byte payload. All Internet hosts are required to accept TCP segments of 536 + 20 = 556 bytes.

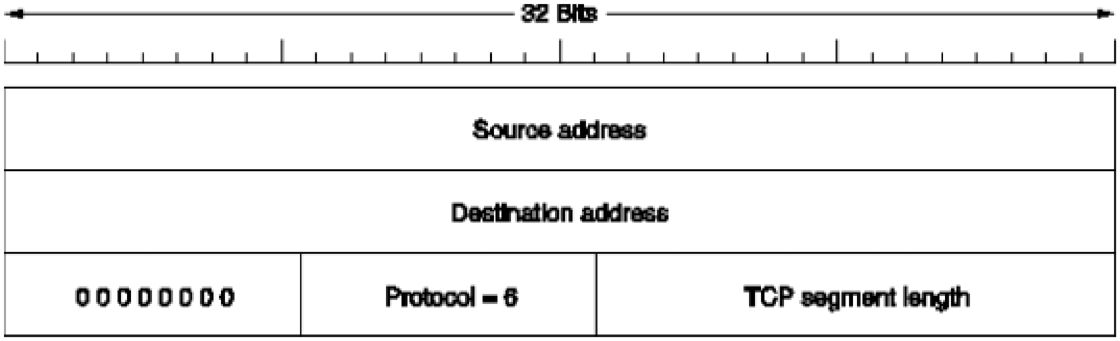


Fig. 6-25. The pseudoheader included in the TCP checksum.

3.2.20. DNS - Domain Name System

Users prefer to refer to hosts, mailboxes, and other resources not by their binary network addresses but using some ASCII strings, such as tana@art.ucsb.edu. Nevertheless, the network itself only understands binary addresses, so some mechanism is required to convert the ASCII string to network addresses. Below we will describe how this mapping is accomplished in the Internet.

The mapping is done by DNS (the Domain Name System).

The essence of DNS is a hierarchical, domain-based naming scheme and a distributed database system for implementing this naming scheme. It is primarily used for mapping host names and email destinations to IP addresses but can also be used for other purposes. DNS is defined in RFC 1034 and 1035.

The basic scheme of the use of DNS is the following: To map a name onto an IP address, an application program calls a library procedure called the resolver, passing it the name as a parameter. The resolver sends a UDP packet to a local DNS server, which then looks up the name and returns the IP address to the resolver, which then returns it to caller. Armed with the IP address, the program then establish a TCP connection with the destination, or send it UDP packets.

3.2.21. The DNS Name Space

Conceptually, the Internet is divided into several hundred top-level *domains*, where each domain covers many hosts. Each domain is partitioned into subdomains, and these are further partitioned, and so on. All these domains can be represented by a tree as in Fig. 7-25. The leaves of the tree represent domains that have no subdomains. A leaf domain may contain a single host, or it may represent a company and contains thousands of hosts.

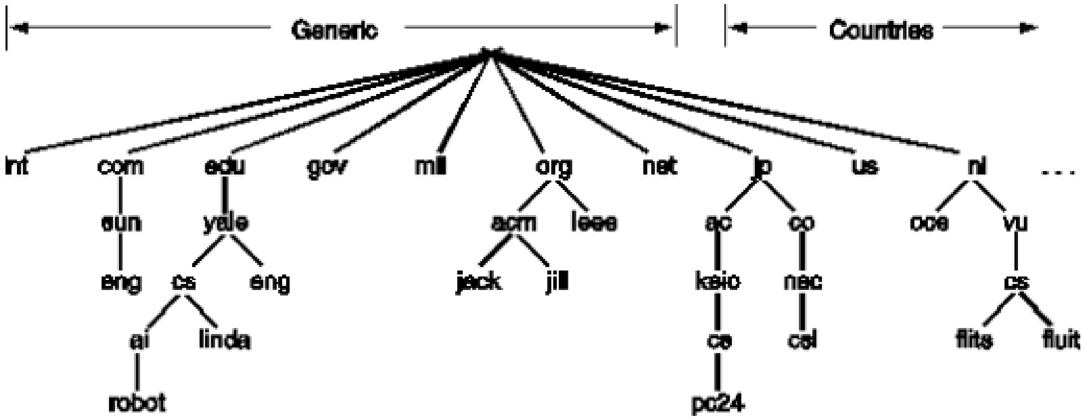


Fig. 7-25. A portion of the Internet domain name space.

The top-level domains are of two kinds: generic and countries. The generic domains are com (commercial), edu (educational institutions), gov (the U.S. federal government), int (certain international organizations), mil (the U.S. armed forces), net (network providers), and org (nonprofit organizations).The country domains include one entry for every country, as defined in ISO 3166.

Each domain is named by the path upward from it to the (unnamed) root. The components are separated by periods (pronounced "dot"). Thus the Faculty of Mathematics and Physics of Comenius University is fmph.uniba.sk.

Domain names are case insensitive, so edu and EDU mean the same thing. Component names can be up to 63 characters long, and full path name must not exceed 255 characters.

In principle, domains can be inserted into the tree in two different ways. For example, cs.yale.edu could equally well be listed under the us country domain as cs.yale.ct.us. In practice, however, nearly all organizations in the U.S. are under a generic domain, and nearly all outside the U.S. are under the domain of their country. There is no rule against registering under two top-level domains, but doing so might be confusing, so few organizations do it.

Each domain controls how it allocates the domains under it. To create a new domain, permission is required of the domain in which it will be included. In this way, name conflicts are avoided. Once a new domain has been created and registered, it can create subdomains without getting permission from anybody higher up the tree.

Naming follows organizational boundaries, not physical networks.

3.2.22. Resource Records

Every domain can have a set of resource records associated with it. For a single host, the most common record is just its IP address, but many other kinds of resource records also exist. When a resolver gives a domain name to DNS, what it gets back are the resource records associated with that name. Thus the real function of DNS is to map domain names onto resource records.

A resource record is a five-tuple. Although they are encoded in binary, in most expositions resource records are presented in ASCII text, one line per resource record. The format we will use is as follows:

Domain_name	Time_to_live	Type	Class	Value

The Domain_name tells the domain to which this record applies. Normally, many records exist for each domain. When a query is made about a domain, all the matching records of the type requested are returned.

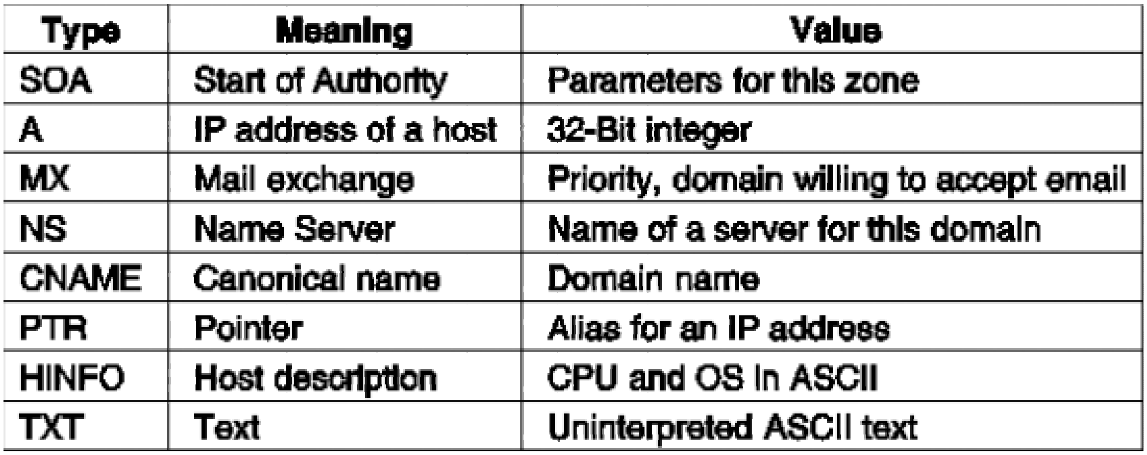
The Time_to_live field gives an indication of how stable the record is. Information that is highly stable is assigned a large value, such as 86400 (the number of seconds in 1 day). Information that are highly volatile is assigned a small value such as 60 (1 minute).

The Type field tells, what kind of record this is. The most important types are:

- SOA - Start of Authority. Provides the name of the primary source of information about the name server’s zone and some further information about it.
- A - IP address of a host. It is the most important record type. It holds a 32-bit IP address for some host. If a host has more network connections, and so more IP addresses, it has a resource record for each of them.
- MX - Mail exchange. It specifies the name of the host prepared to accept email for the specified domain.
- NS - Name server. It specifies name servers. For example, every DNS database normally has an NS record for each of the top-level domains.
- CNAME - Canonical name. This record allows aliases to be created.
- PTR - Pointer. This is an allias for an IP address. Records of this type are nearly always used to associate a name with an IP address to allow lookups of the IP address and return the name of the corresponding machine. We omit the details of this process here.
- HINFO - Host description. This record allow people to find out what kind of machine and operating system a domain corresponds to.
- TXT - Text. This record allows domains to identify themselves in arbitrary ways.

The Class field is always equal IN for Internet information. For non-Internet information, other codes can be used.

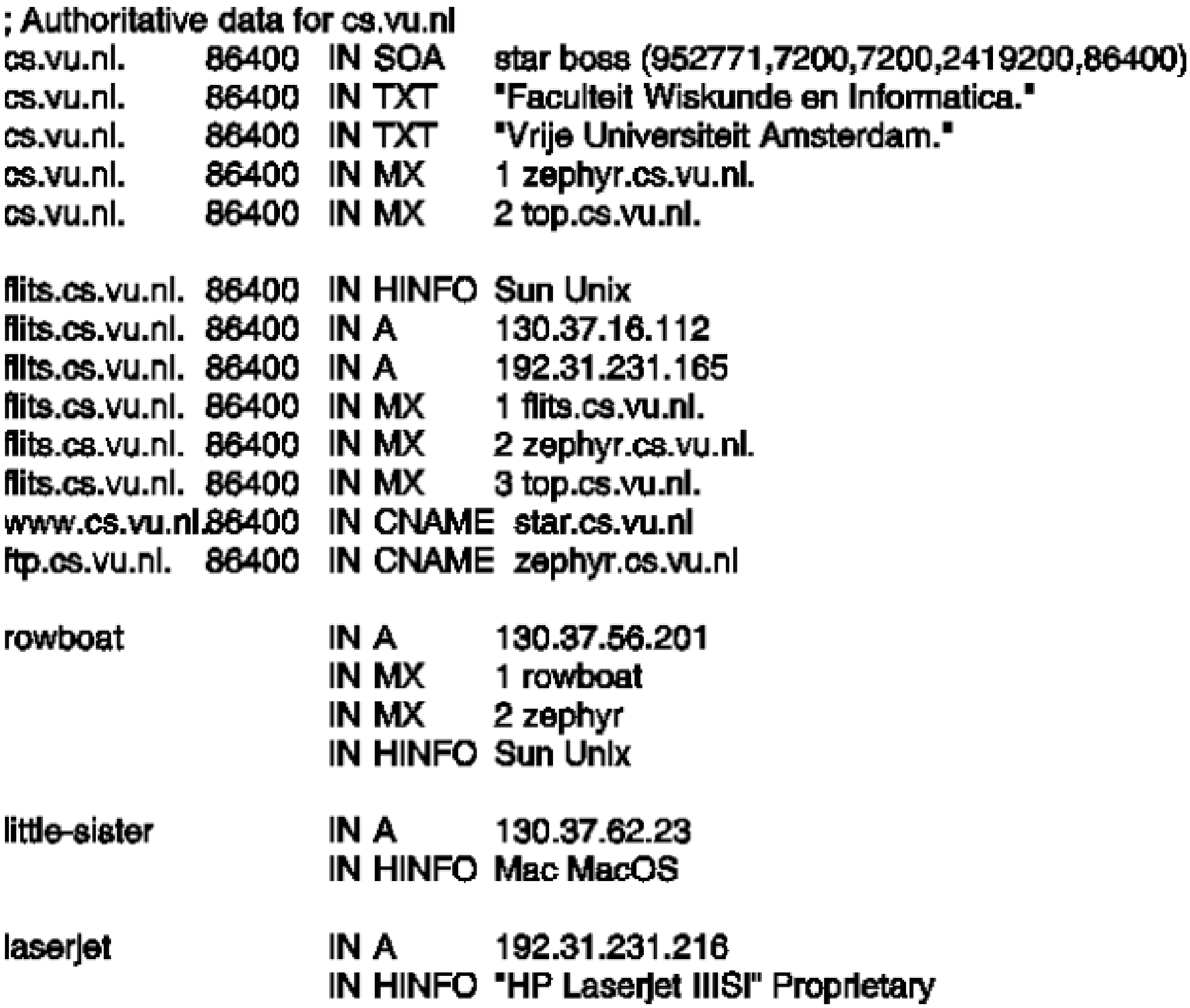
The Value field can contain a number, a domain name, or an ASCII string. The semantics depends on the record type. A short description of the Value fields for each of the principal record types is given in Fig. 7-26.



Type	Meaning	Value
SOA	Start of Authority	Parameters for this zone
A	IP address of a host	32-Bit integer
MX	Mail exchange	Priority, domain willing to accept email
NS	Name Server	Name of a server for this domain
CNAME	Canonical name	Domain name
PTR	Pointer	Alias for an IP address
HINFO	Host description	CPU and OS in ASCII
TXT	Text	Uninterpreted ASCII text

Fig. 7-26. The principal DNS resource record types.

As an example of the kind of information one can find in the DNS database of a domain, see Fig. 7-27. This figure depicts part of a database for the cs.vu.nl domain shown in Fig. 7-25. The database contains seven types of resource records.



```
; Authoritative data for cs.vu.nl
cs.vu.nl.        86400   IN SOA     star boss (952771,7200,7200,2419200,86400)
cs.vu.nl.        86400   IN TXT     "Faculteit Wiskunde en Informatica."
cs.vu.nl.        86400   IN TXT     "Vrije Universiteit Amsterdam."
cs.vu.nl.        86400   IN MX      1 zephyr.cs.vu.nl.
cs.vu.nl.        86400   IN MX      2 top.cs.vu.nl.

flits.cs.vu.nl.  86400   IN HINFO   Sun Unix
flits.cs.vu.nl.  86400   IN A       130.37.16.112
flits.cs.vu.nl.  86400   IN A       192.31.231.165
flits.cs.vu.nl.  86400   IN MX      1 flits.cs.vu.nl.
flits.cs.vu.nl.  86400   IN MX      2 zephyr.cs.vu.nl.
flits.cs.vu.nl.  86400   IN MX      3 top.cs.vu.nl.
www.cs.vu.nl86400        IN CNAME    star.cs.vu.nl
ftp.cs.vu.nl.    86400   IN CNAME    zephyr.cs.vu.nl

rowboat                  IN A       130.37.56.201
                         IN MX      1 rowboat
                         IN MX      2 zephyr
                         IN HINFO   Sun Unix

little-sister            IN A       130.37.62.23
                         IN HINFO   Mac MacOS

laserjet                 IN A       192.31.231.216
                         IN HINFO   "HP Laserjet IIISi" Proprietary
```

Fig. 7-27. A portion of a possible DNS database for cs.vu.nl

The first noncomment line of Fig. 7-27 gives some basic information about the domain, which will not concern us further.

The next two lines give textual information about where the domain is located.

Then come two entries giving the first and second places to try to deliver email sent to person@cs.vu.nl. The zephyr (a specific machine) should be tried first. If that fails, the top should be tried next.

Next 3 lines tell that the flits is a Sun workstation running UNIX and give both of its IP addresses.

Further three lines give choices for handling email sent to flits.cs.vu.nl.

Next comes an alias, www.cs.vu.nl, so this address can be used without designating a specific machine. Similarly ftp.cs.vu.nl.

The next four lines contain a typical entry for a workstation, in this case rowboat.cs.vu.nl. The information provided contains the IP address, the primary and secondary mail drops, and information about the machine. Then comes an entry for a non-UNIX system that is not capable of receiving mail itself, followed by an entry for a laser printer.

IP addresses for root servers needed to look up distant hosts are not in this file. They are present in a system configuration file loaded into the DNS cache when the server is booted. They have very long timeouts, so once loaded, they are never purged from the cache.

3.2.23. Name Servers

In practice, one single name server cannot contain the entire DNS database. So the DNS name space is divided up into nonoverlapping zones and each zone contains name servers holding the authoritative information about that zone (See Fig. 7-28 as a possible way how to divide up the name space from Fig. 7-25). Normally, a zone will have one primary name server, which gets its information from a file on its disk, and one or more secondary name servers, which get their information from the primary name server.

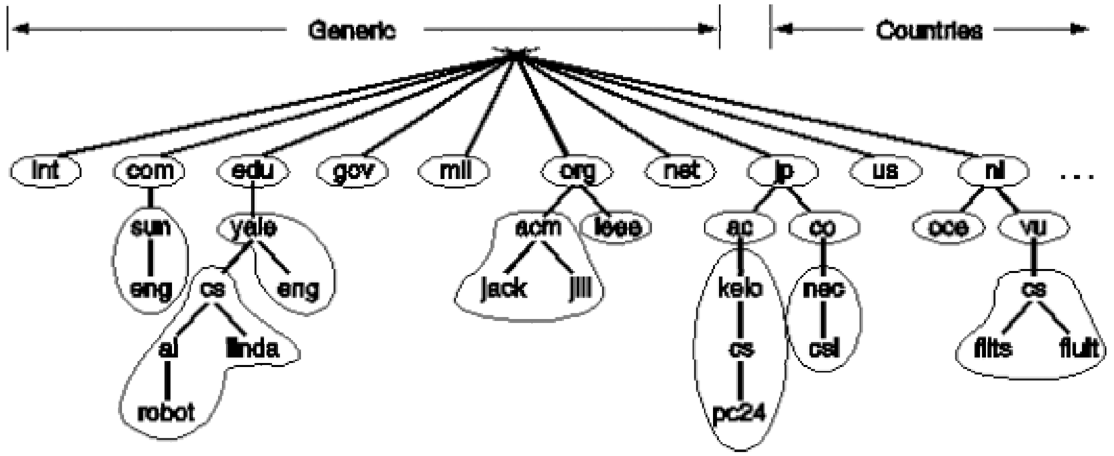


Fig. 7-28. Part of the DNS name space showing the division into zones.

When a resolver has a query about a domain name, it passes the query to one of the local name servers. If the domain being sought falls under the jurisdiction of the name server, it returns the authoritative resource records. An authoritative record is one that comes from the authority that manages the record, and thus is always correct. Authoritative records are in contrast with cached records, which may be out of date.

If, however, the domain is remote and no information about the requested domain is available locally, the name server sends a query message to the top-level name server for the domain requested. If it also does not know the answer, it sends it to one of its children, and so on. When a server with the authoritative resource record is encountered, the response is sent back through single name servers in the chain. For an example, see Fig. 7-29, here the IP address of the host linda.cs.yale.edu was sought by the resolver on flits.cs.vu.nl.

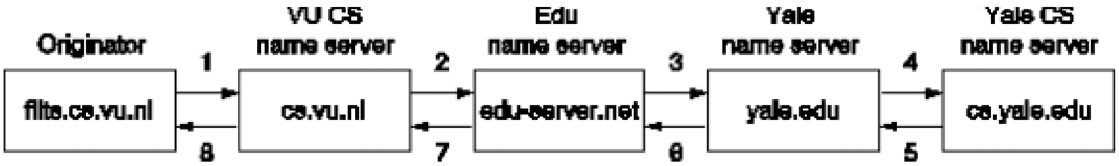


Fig. 7-29. How a resolver looks up a remote name in eight steps.

Once the record get back to the name server cs.vu.nl, it will be entered into a cache there, in case it is needed later. However, this information is not authoritative, so it should not live too long. This is the reason that the Time_to_live field is included in each resource record. It tells remote name servers how long to cache records.

The query method described here is known as a *recursive query*. An alternative form is also possible. In this form, when a query cannot be satisfied locally, the query fails, but the name of the next server on the line to try is returned. This procedure gives the client more control over the search process.

When a DNS client fails to get a response before its timer goes off, it normally will try another server next time.

3.3. The Network Layer in ATM Networks

The layers of the ATM model (Fig. 1-30) do not map onto the OSI layers especially well, which leads to ambiguities. The OSI data link layer deals with framing and transfer protocols between two machines on the same physical wire (or fiber). Data link protocols are single-hop protocols. They do not deal with end-to-end connections because switching and routing do not occur in the data link layer.

The lowest layer that goes from source to destination, and thus involves routing and switching (i.e., is multihop), is the network layer. The ATM layer deals with moving cells from source to destination and definitely involves routing algorithms and protocols within the ATM switches. It also deals with global addressing. Thus functionally, the ATM layer performs the work expected of the network layer.

Confusion arises because many people in the ATM community regard the ATM layer as a data link layer, or when doing LAN emulation, even physical layer. Many people in the Internet community also regard it as a data link layer because they want to put IP on top of it, and making the ATM layer a data link layer fits well with this idea.

But due to its characteristics, ATM layer is a network layer.

The ATM layer is connection oriented, both in terms of the service it offers and the way it operates internally. The basic element of the ATM layer is the virtual circuit (officially called a virtual channel). A virtual circuit is normally a connection from one source to one destination. Virtual circuits are unidirectional, but a pair of circuits can be created at the same time. Both parts of the pair are addressed by the same identifier, so effectively a virtual circuit is full duplex. However, the channel capacity and other properties may be different in the two directions and may be even zero for one of them.

The ATM layer does not provide any acknowledgments, it leaves error control to higher layers. The reason for this design is that ATM was designed for use on fiber optics networks, which are highly reliable. Furthermore, ATM networks are often used for real-time traffic, such as audio and video. For this kind of traffic, retransmitting an occasional bad cell is worse than just ignoring it.

Despite its lack of acknowledgment, the ATM layer does provide one hard guarantee: cells sent along a virtual circuit will never arrive out of order. The ATM subnet is permitted to discard cells if congestion occurs but under no conditions may it reorder the cells sent on a single virtual circuit.

The ATM layer supports a two-level connection hierarchy that is visible to the transport layer. Along any transmission path from a given source to a given destination, a group of virtual circuits can be grouped together into what is called a virtual path (Fig. 5-61). Conceptually, a virtual path is like a bundle of twisted copper pairs: when it is rerouted, all the pairs (virtual circuits) are rerouted together.

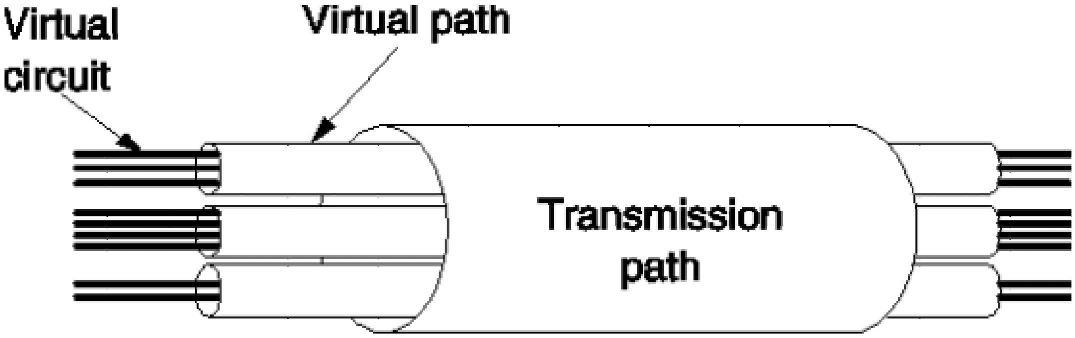


Fig. 5-61. A transmission path can hold multiple virtual paths, each of which can hold multiple virtual circuits.

3.3.1. Cell Formats

In the ATM layer, two *interfaces* are distinguished:

- UNI (User-Network Interface) - defines the boundary between a host and an ATM network.
- NNI (Network-Network Interface) - applies to the line between two ATM switches (ATM switch is the ATM term for router).

In both cases the cells consist of a 5-byte header followed by a 48-byte payload, but the two headers are slightly different. Cells are transmitted leftmost byte first and leftmost bit within a byte first.

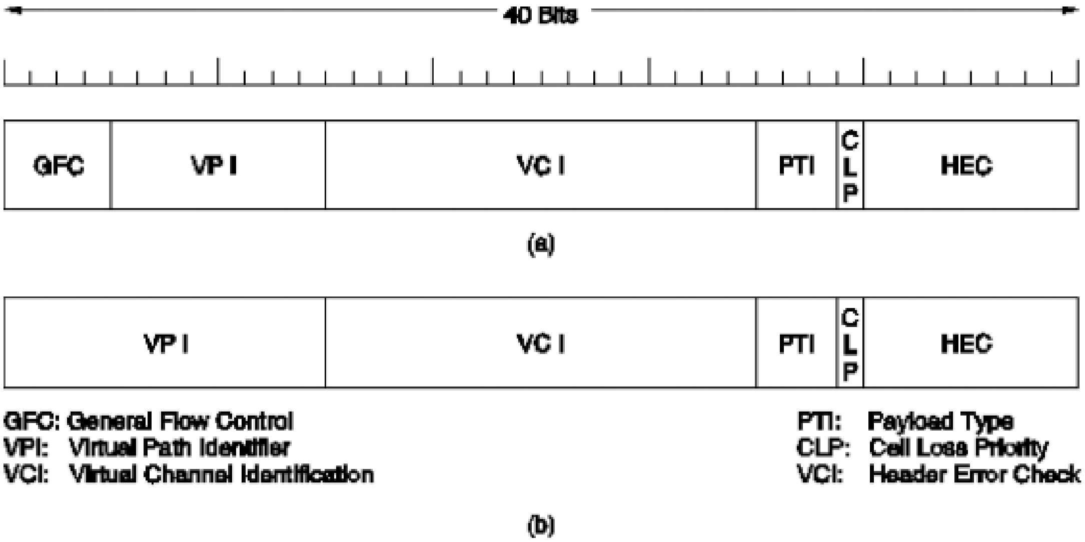
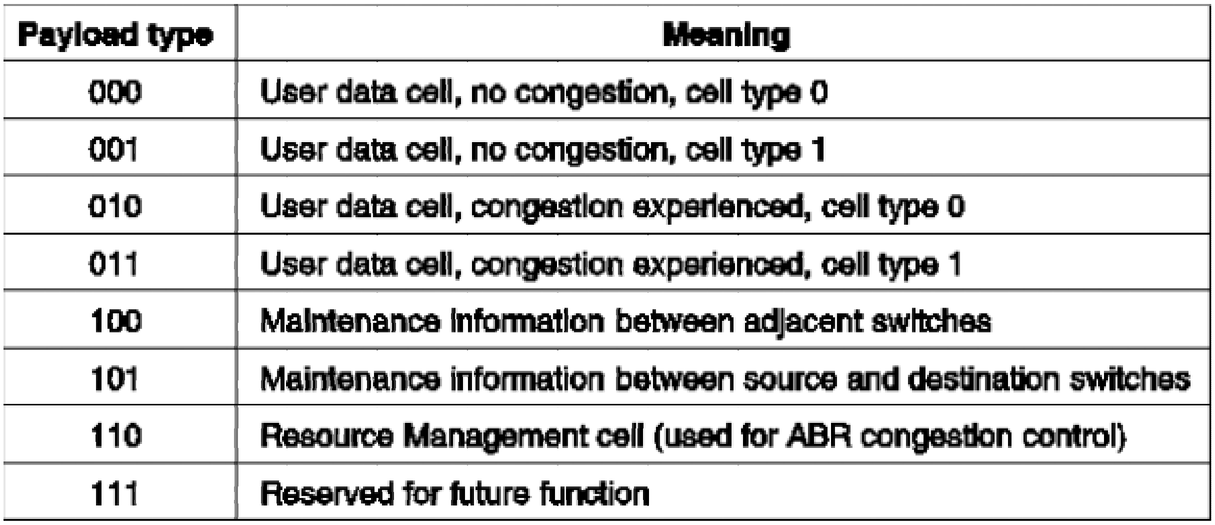


Fig. 5-62. (a) The ATM layer header at the UNI. (b) The ATM layer header at the NNI.

The meaning of single fields in the headers (Fig. 5-62):

- VPI (Virtual Path Identifier) - occupies bits 0 - 11. This field contains an integer selecting a particular virtual path. In cells between a host and a network (UNI), VPI occupies in fact just bits 4 - 11, bits 0 - 3 are called GFC (General Flow Control) field. But this field has no significance and the network ignores it. It was originally conceived as perhaps having some utility for flow control or priority between hosts and networks.
- VCI (Virtual Channel Identification) - occupies bits 12 - 27. This field selects a particular virtual circuit within the chosen virtual path.
- PTI (Payload Type) - occupies bits 28 - 30. This field defines the type of payload the cell contains (Fig. 5-63). Cell type information are provided by users, congestion information are network supplied. In other words, a cell sent with PTI 000 might arrive with 010 to warn destination of problems underway.



Payload type	Meaning
000	User data cell, no congestion, cell type 0
001	User data cell, no congestion, cell type 1
010	User data cell, congestion experienced, cell type 0
011	User data cell, congestion experienced, cell type 1
100	Maintenance information between adjacent switches
101	Maintenance information between source and destination switches
110	Resource Management cell (used for ABR congestion control)
111	Reserved for future function

Fig. 5-63. Values of the PTI field.

- CLP (Cell Loss Priority) - occupies bit 31. This bit is set by a host to differentiate between high-priority traffic and low-priority traffic. If congestion occurs and cells must be discarded, switches first attempt to discard cells with CLP set to 1 before throwing out any set to 0.

- HEC (Header Error Check) - occupies bits 32 - 39. This field is a checksum over the header. It does not check the payload. The chosen code can correct all single-bit errors and detect about 90% of all multibit errors. Various studies have shown that the vast majority of errors on optical links are single-bit errors.

Following the header comes 48 bytes of payload. Not all 48 bytes are available to the user, however, since some of the AAL protocols put their headers and trailers inside the payload.

So the NNI format is the same as the UNI format, except that the GFC field is not present and those 4 bits are used to make VPI field 12 bits instead of 8.

3.3.2. Connection Setup

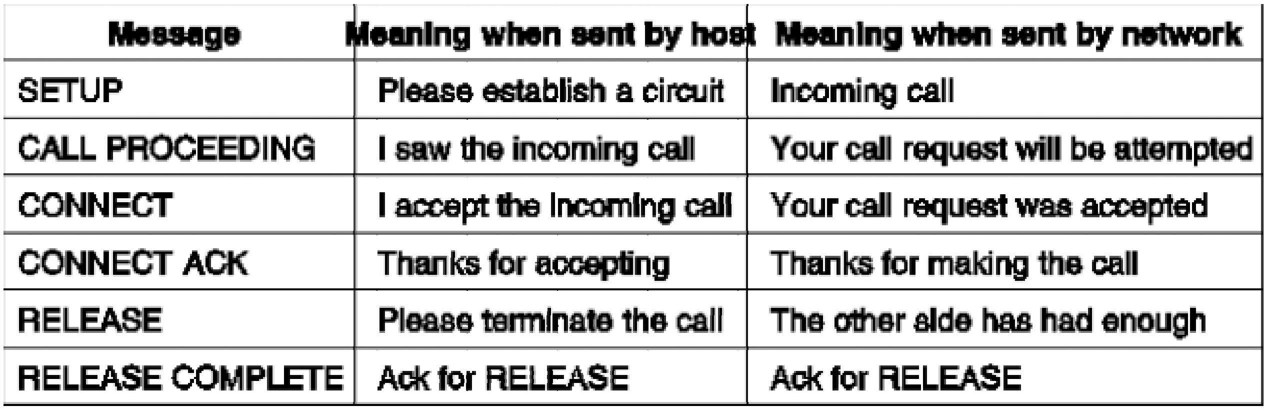
ATM supports both *permanent virtual circuits* (i.e., always present, like leased lines) and *switched virtual circuits* (they have to be established each time they are used, like making phone calls).

We will describe how switched virtual circuits are established.

Technically, connection setup is not part of the ATM layer but is handled by the control plane (Fig. 1-30) using a highly-complex ITU protocol called Q.2931.

Several ways are provided for setting up a connection. The normal way is to first acquire a virtual circuit for signaling and use it. To establish such a circuit, cell containing a request are sent on virtual path 0, virtual circuit 5. If successful, a new virtual circuit is opened on which connection setup requests and replies can be sent and received.

Virtual circuit establishment uses the six message types (Fig. 5-64). Each message occupies one or more cells and contains the message type, length, and parameters. The messages can be sent by a host to the network or by the network to a host. Various other status and reporting messages also exist but are not mentioned here.



Message	Meaning when sent by host	Meaning when sent by network
SETUP	Please establish a circuit	Incoming call
CALL PROCEEDING	I saw the incoming call	Your call request will be attempted
CONNECT	I accept the incoming call	Your call request was accepted
CONNECT ACK	Thanks for accepting	Thanks for making the call
RELEASE	Please terminate the call	The other side has had enough
RELEASE COMPLETE	Ack for RELEASE	Ack for RELEASE

Fig. 5-64. Messages used for connection establishment and release.

The normal procedure for establishing a call is for a host to send a SETUP message on a special virtual circuit. The network then responds with CALL PROCEEDING to acknowledge receipt of the request. As the SETUP message propagates toward the destination, it is acknowledged at each hop by CALL PROCEEDING.

When the SETUP message finally arrives, the destination host can respond with CONNECT to accept the call. The network then sends a CONNECT ACK message to indicate that it has received the

CONNECT message. As the CONNECT message propagates back toward the originator, each switch receiving it acknowledges it with a CONNECT ACK message (Fig. 5-65).

When a host wants to terminate a virtual circuit, it just sends a RELEASE message that propagates to the other end and causes the circuit to be released. Each hop along the way, the message is acknowledged (Fig. 5-65).

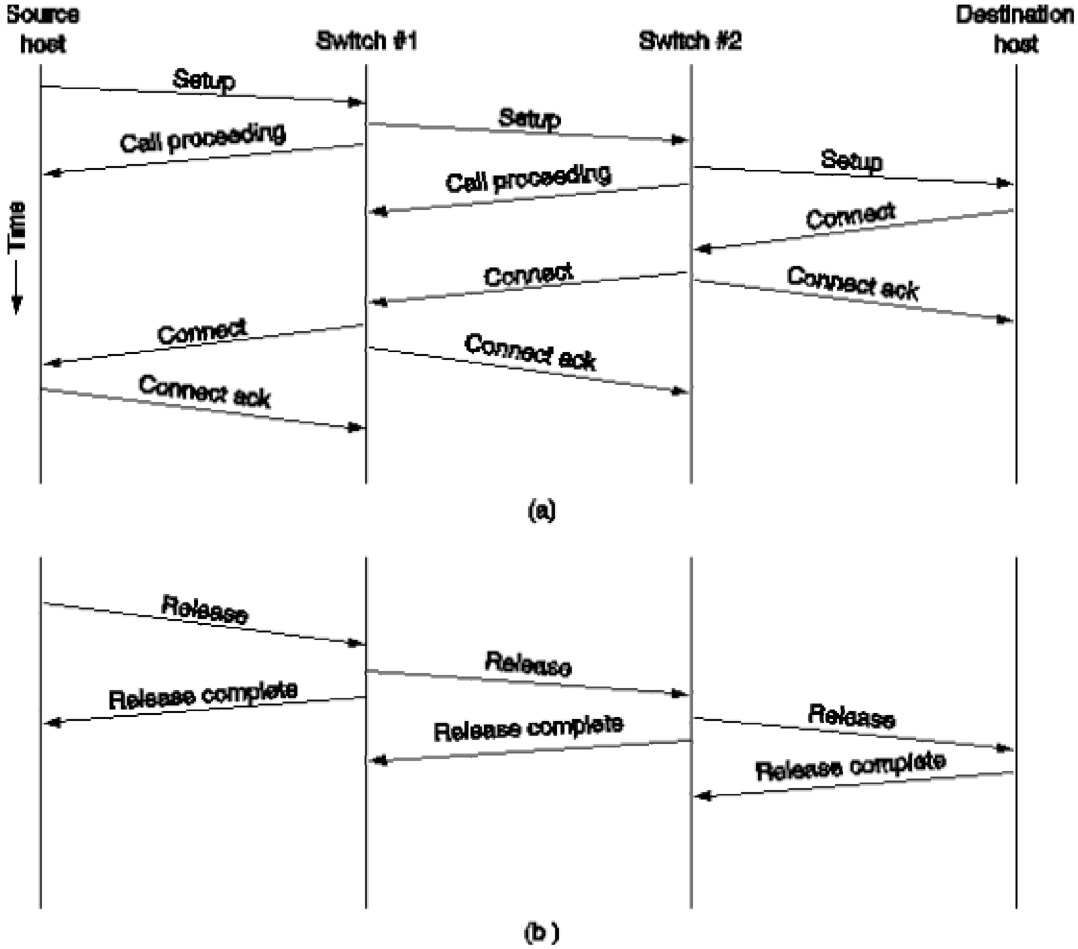


Fig. 5-65. (a) Connection setup in an ATM network. (b) Connection release.

ATM networks allow multicast channels to be set up. A multicast channel has one sender and more than one receiver. These are constructed by setting up a connection to one of the destinations in the usual way. Then ADD PARTY message is sent to attach a second destination to the virtual circuit returned by the previous call. Additional ADD PARTY messages can be sent afterwards to increase the size of the multicast group.

Each SETUP message must specify an address of the destination. ATM addresses come in three forms.

The first is 20 bytes long and is based on OSI addresses. The first byte indicates which of the three formats the address is in. In the first format, bytes 2 and 3 specify a country, and byte 4 gives the format of the rest of the address, which contains a 3-byte authority, a 2-byte domain, a 2-byte area, and a 6-byte address, plus some other items. In the second format, bytes 2 and 3 designate an international organization instead of a country. The rest of the address is the same as in format 1. Alternatively, an older form of addressing (CCITT E.164) using 15-digit decimal ISDN telephone number is also permitted.

3.3.3. Routing and Switching

The ATM standard does not specify any particular routing algorithm, so the carrier is free to chose among the different algorithms.

ATM layer routing is based on the VPI field, not on the VCI field, except at the final hop in each direction, when cells are sent between a switch and a host. Between two switches, only the virtual path is used.

Let us now see how cells could be routed within an interior switch (one that is attached only to other switches and not to hosts). Let us consider a concrete example, the Omaha switch in Fig. 5-66. For each of its five incoming lines, it has a table, *vpi_table*, indexed by incoming VPI that tells which of five outgoing lines to use and what VPI to put in outgoing cells. Let us assume the five lines are numbered from 0 to 4, as in the figure. For each outgoing line, the switch maintains a bit map telling which VPIs are currently in use on that line.

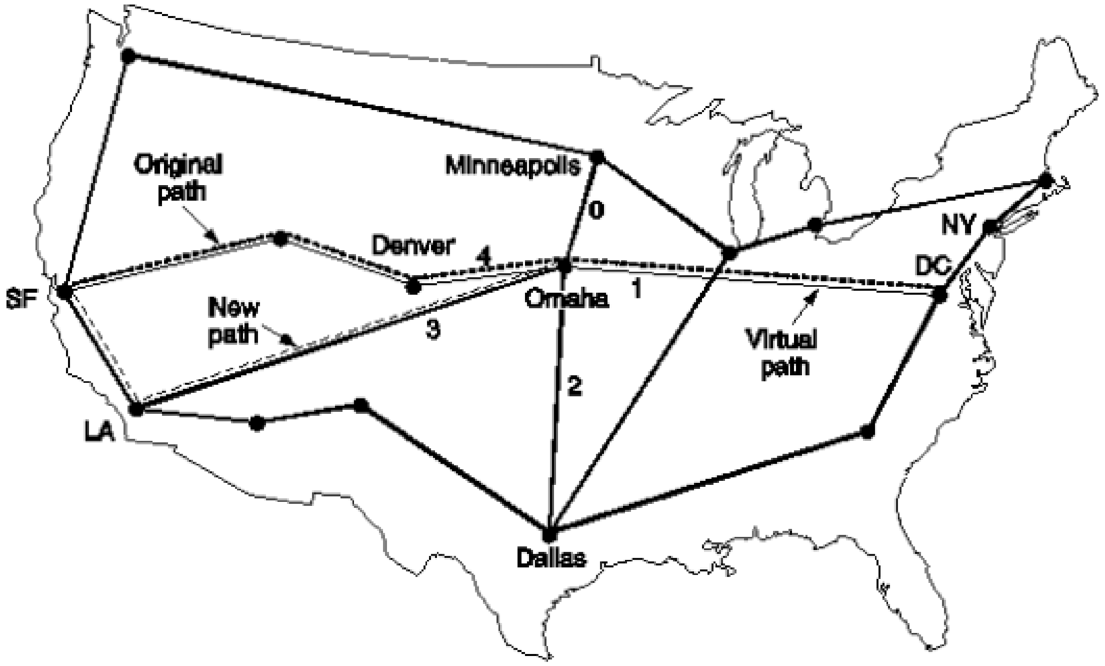


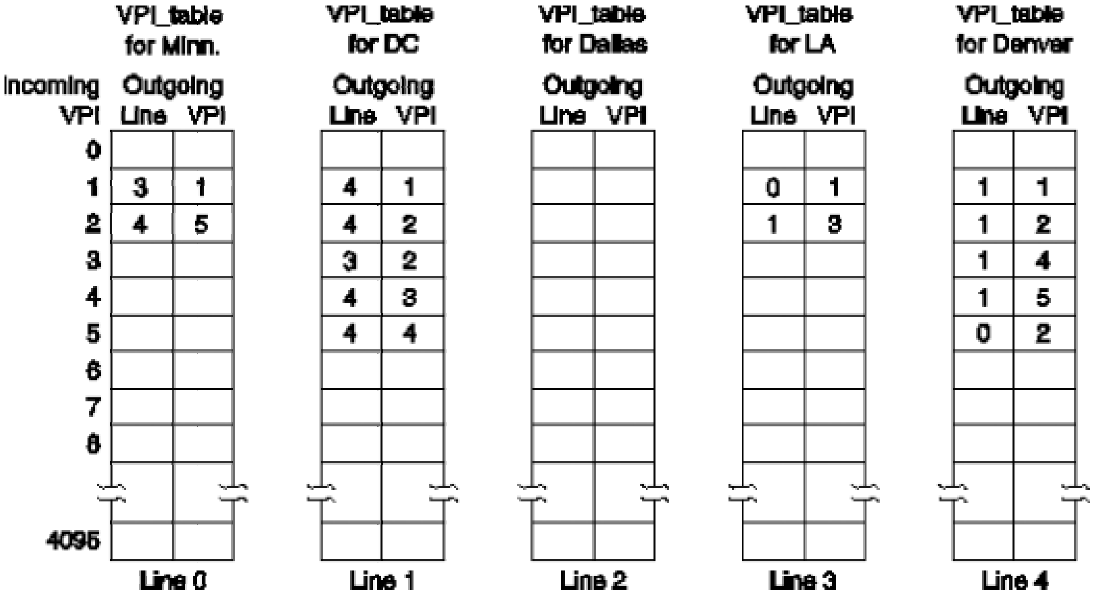
Fig. 5-66. Rerouting a virtual path reroutes all of its virtual circuits.

When the switch is booted, all the entries in all vpi_table structures are marked as not in use. Similarly, all the bit maps are marked to indicate that all VPIs are available (except the reserved ones). Now suppose calls come as shown in Fig. 5-67.

Source	Incoming line	Incoming VPI	Destination	Outgoing line	Outgoing VPI	Path:
NY	1	1	SF	4	1	New
NY	1	2	Denver	4	2	New
LA	3	1	Minneapolis	0	1	New
DC	1	3	LA	3	2	New
NY	1	1	SF	4	1	Old
SF	4	3	DC	1	4	New
DC	1	5	SF	4	4	New
NY	1	2	Denver	4	2	Old
SF	4	5	Minneapolis	0	2	New
NY	1	1	SF	4	1	Old

Fig. 5-67. Some routes through the Omaha switch of Fig. 5-66.

As each virtual path (and virtual circuit) is set up, entries are made in the tables. We will assume the virtual circuits are full duplex, so that each one setup results in two entries, one for the forward traffic from the source and one for the reverse traffic from the destination.



	VPI_table for Minn.		VPI_table for DC		VPI_table for Dallas		VPI_table for LA		VPI_table for Denver	
Incoming	Outgoing		Outgoing		Outgoing		Outgoing		Outgoing	
VPI	Line	VPI	Line	VPI	Line	VPI	Line	VPI	Line	VPI
0										
1	3	1	4	1			0	1	1	1
2	4	5	4	2			1	3	1	2
3			3	2					1	4
4			4	3					1	5
5			4	4					0	2
6										
7										
8										
4095										
	Line 0		Line 1		Line 2		Line 3		Line 4	

Fig. 5-68. Table entries for the routes of Fig. 5-66.

The tables corresponding to the routes of Fig. 5-67 are shown in Fig. 5-68. For example, the first call generates the (4,1) entry for VPI 1in the DC table because it refers to cells coming in on line 1 with VPI 1 and going to SF. However, an entry is also made in the Denver table for VPI 1 showing that cells coming in from Denver with VPI 1 should go out on line 1with VPI 1. These are cells traveling the other way (from SF to NY) on this virtual path. Note that in some cases two or three virtual circuits are sharing a common path. No new table entries are needed for additional virtual circuits connecting a source and destination that already have a path assigned.

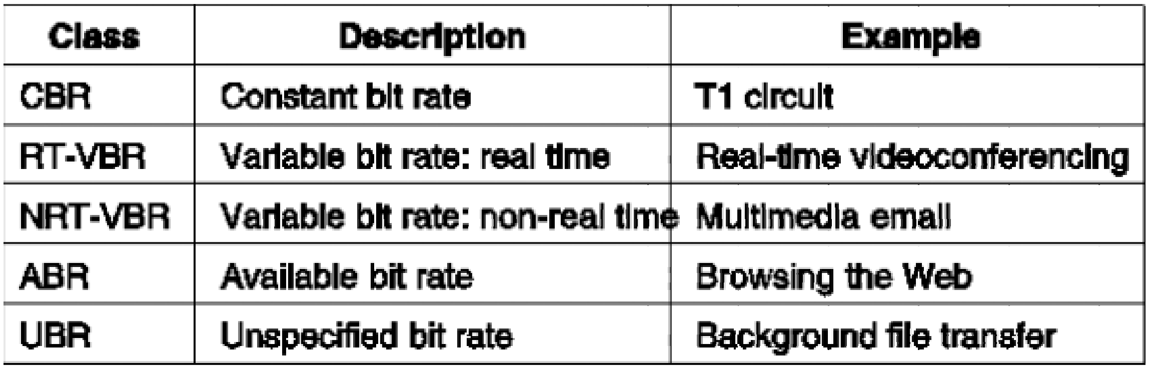
Now we can explain how cells are processed inside a switch. Suppose that a cell arrives on line 1 (DC) with VPI 3. The switch hardware or software uses the 3 as an index into the table for line 1 and sees that the cell should go out on line 3 (LA) with VPI 2. It overwrites the VPI field with a 2 and the switch gets the cell from its current input buffer to line 3.

At this point it is straightforward to see how an entire bundle of virtual circuits can be rerouted, as is done in Fig. 5-66. By changing the entry for VPI 1 in the DC table from (4,1) to (3,3), cells from NY headed for SF will be diverted to LA. Of course, the LA switch has to be informed of this event, so the switch has to generate and send a SETUP message to LA to establish the new path with VPI 3.Once this path has been set up, all the virtual circuits from NY to SF are now rerouted via LA, even if there are thousands of them. If virtual paths did not exist, each virtual circuit would have its own table entry and would have to be rerouted separately.

The discussion above is about ATM in WANs. In a LAN, matters are much simpler.

3.3.4. Service Categories

ATM networks offers the following service categories:



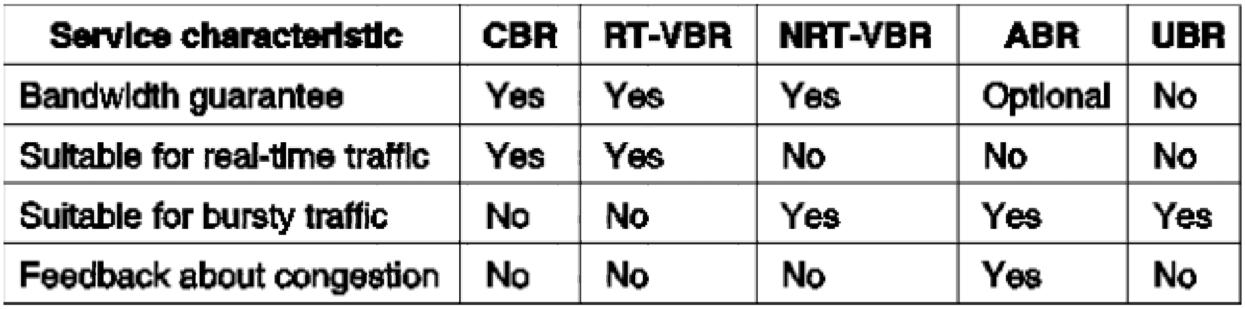
Class	Description	Example
CBR	Constant bit rate	T1 circuit
RT-VBR	Variable bit rate: real time	Real-time videoconferencing
NRT-VBR	Variable bit rate: non-real time	Multimedia email
ABR	Available bit rate	Browsing the Web
UBR	Unspecified bit rate	Background file transfer

Fig. 5-69. The ATM service categories.

- CBR (Constant Bit Rate) - this class is intended to emulate a copper wire or optical fiber. Bits are put on one end and they come off the other end. No error checking, flow control, or other processing is done. With CBR class a smooth transition between the current telephone system and future B-ISDN systems will be made since voice-grade PCM channels, T1 circuits, and most of the rest of the telephone system use constant rate.
- VBR (Variable Bit Rate) - is divided into two subclasses, for real-time and non-real time. RT-VBR is intended for services that have variable bit rates combined with stringent real-time requirements (e.g. interactive compressed video). Non-real time VBR subclass is for traffic where timely delivery is important but a certain amount of jitter can be tolerated by application (e.g. multimedia email that is typically spooled to the receiver’s local disk before being displayed).
- ABR (Available Bit Rate) - is designed for bursty traffic whose bandwidth range is known roughly. With ABR it is possible to say, for example, that the capacity between two points always be 5Mbps, but might have peaks up to 10 Mbps. The system will then guarantee 5Mbps all the time, and do its best to provide 10 Mbps when needed but with no promises. ABR is the only service category in which the network provides rate feedback to the sender, asking it to slow down when congestion occurs.
- UBR (Unspecified Bit Rate) - makes no promises. This category is well suited for sending IP packets, since IP also makes no promises about delivery. All UBR cells are accepted, and if

there is capacity left over, they will also be delivered. If congestion occurs, UBR cells will be discarded, with no feedback to the sender and no expectation that the sender slows down. UBR is attractive because it is cheaper than other classes. It can be used by applications that have no delivery constraints and want to do their own error control and flow control (e.g. file transfer, email).

The properties of various service categories are summarized in Fig. 5-70.

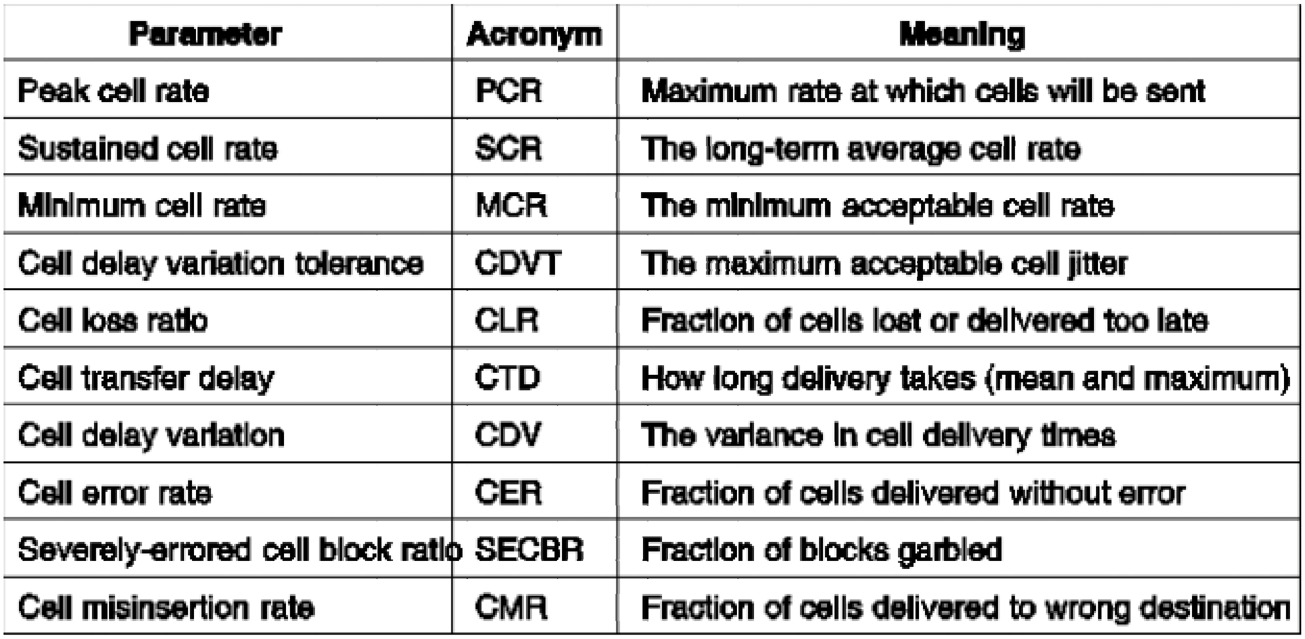


Service characteristic	CBR	RT-VBR	NRT-VBR	ABR	UBR
Bandwidth guarantee	Yes	Yes	Yes	Optional	No
Suitable for real-time traffic	Yes	Yes	No	No	No
Suitable for bursty traffic	No	No	Yes	Yes	Yes
Feedback about congestion	No	No	No	Yes	No

Fig. 5-70. Characteristics of the ATM service categories.

3.3.5. Quality of Service

Quality of service is an important issue for ATM networks, in part because they are used for real-time traffic such as audio and video. The customer and the network operator or carrier must agree on a contract defining the service.



Parameter	Acronym	Meaning
Peak cell rate	PCR	Maximum rate at which cells will be sent
Sustained cell rate	SCR	The long-term average cell rate
Minimum cell rate	MCR	The minimum acceptable cell rate
Cell delay variation tolerance	CDVT	The maximum acceptable cell jitter
Cell loss ratio	CLR	Fraction of cells lost or delivered too late
Cell transfer delay	CTD	How long delivery takes (mean and maximum)
Cell delay variation	CDV	The variance in cell delivery times
Cell error rate	CER	Fraction of cells delivered without error
Severely-errored cell block ratio	SECBR	Fraction of blocks garbled
Cell misinsertion rate	CMR	Fraction of cells delivered to wrong destination

Fig. 5-71. Some of the quality of service parameters.

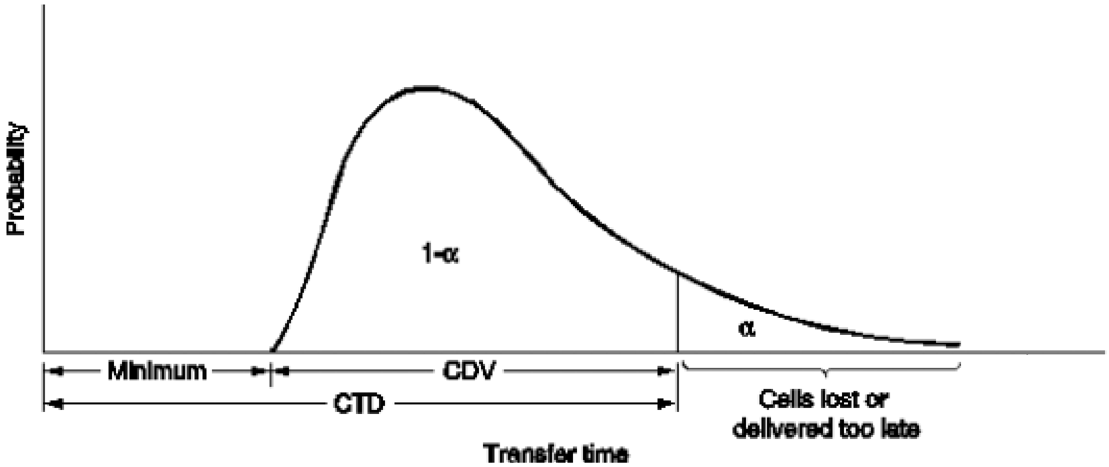


Fig. 5-72. The probability density function for cell arrival times.

The contract between the customer and the carrier has three parts:

1. Traffic to be offered
2. The service agreed upon.
3. The compliance requirements.

The contract may be different for each direction. For a video-on-demand application, the required bandwidth from the user's remote control to the video server might be 1200 bps. In the other direction it might be 5Mbps.

To make it possible to have concrete traffic contracts, the ATM standard defines a number of QoS (Quality of Service) parameters whose values the customer and carrier can negotiate. The most important ones are the following:

The first three parameters specify how fast the user wants to send.

- PCR (Peak Cell Rate) - the maximum rate at which the sender is planning to send cells.
- SCR (Sustained Cell Rate) - expected or required cell rate averaged over a long time interval. For CBR traffic, SCR will be equal to PCR. The PCR/SCR ratio is one measure of the burstiness of the traffic.
- MCR (Minimum Cell Rate) - minimum number of cells/sec that the customer considers acceptable. If the customer and carrier agree to setting MCR to 0, then ABR service becomes similar to UBR service.
- CVDT (Cell Variation Delay Tolerance) - tells how much variation will be present in cell transmission times. For a perfect source operating at PCR, every cell will appear exactly 1/PCR after the previous one. For a real source operating at PCR, some variation will occur in cell transmission time. The question is: how much variation is acceptable? CDVT controls the amount of variability acceptable using a leaky bucket algorithm (Fig. 5-73 and Fig. 5-74).

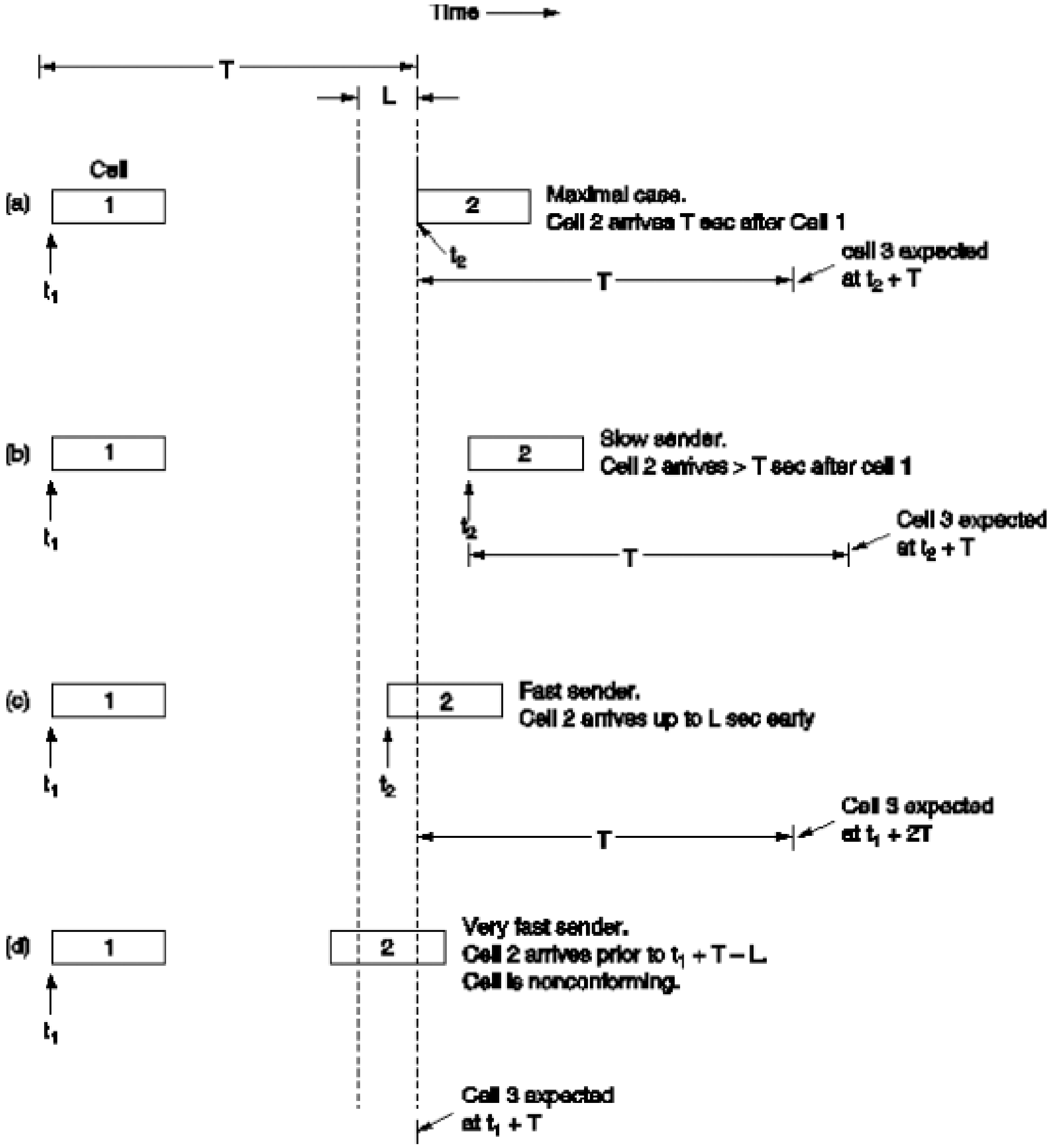

Time
T
L
Cell
(a)
1
2
Maximal case.
Cell 2 arrives T sec after Cell 1
t_2
t_1
T
cell 3 expected at t_2 + T
(b)
1
2
Slow sender.
Cell 2 arrives > T sec after cell 1
t_1
t_2
T
Cell 3 expected at t_2 + T
(c)
1
2
Fast sender.
Cell 2 arrives up to L sec early
t_1
t_2
T
Cell 3 expected at t_1 + 2T
(d)
1
2
Very fast sender.
Cell 2 arrives prior to t_1 + T – L.
Cell is nonconforming.
t_1
Cell 3 expected at t_1 + T

Fig. 5-73. The generic cell rate algorithm.

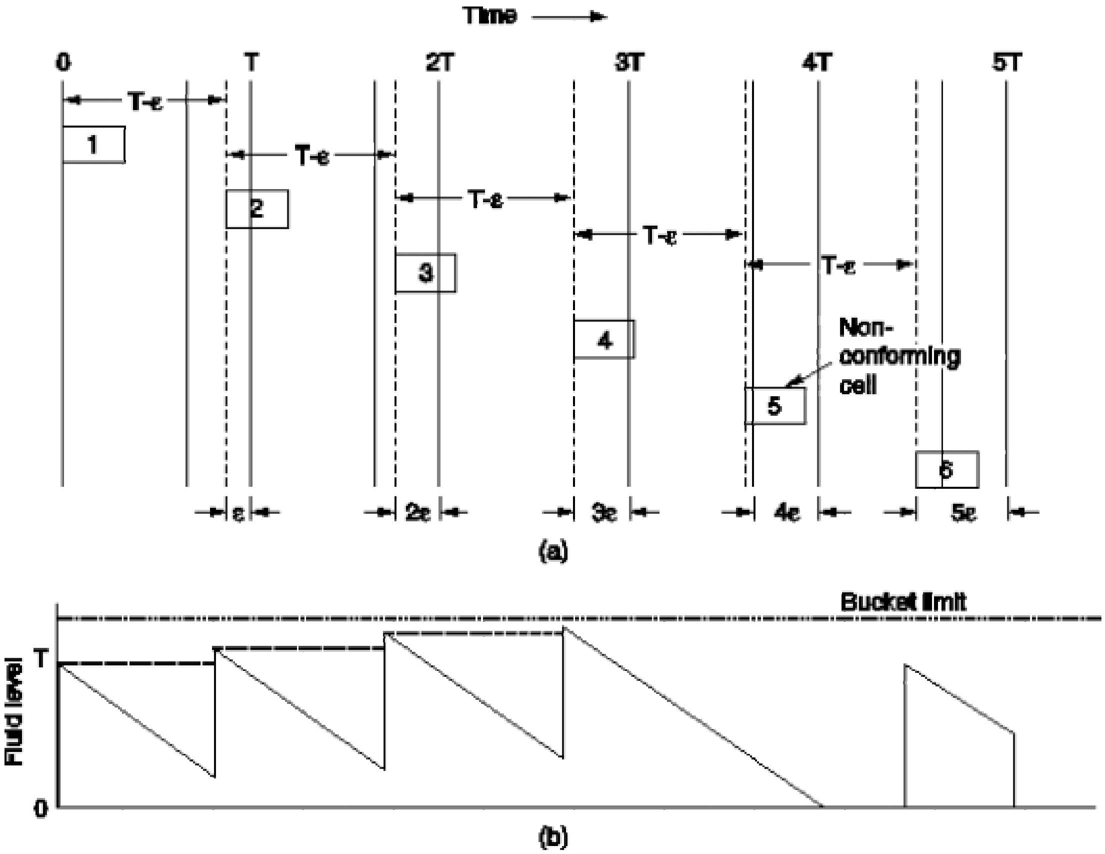


Fig. 5-74. (a) Sender trying to cheat. (b) The same cell arrival pattern, but now viewed in terms of a leaky bucket.

The next three parameters describe characteristics of the network and are measured at the receiver.

- CLR (Cell Loss Ratio) - the fraction of the transmitted cells that are not delivered at all or are delivered so late as to be useless.
- CTD (Cell Transfer Delay) - the average transit time from source to destination.
- CDV (Cell Delay Variation) - measures how uniformly the cells are delivered (Fig. 5-72). By choosing a value of CTD, the customer and the carrier are, in effect, agreeing, on how late a cell can be delivered and still count as a correctly delivered cell. Normally, CDV will be chosen so that, (, the fraction of cells that are rejected for being too late will be on the order of 10-10 or less. CDV measures the spread in arrival times. For real-time traffic, this parameter is often more important than CTD.

The last three parameters specify characteristics of the network. They are generally not negotiable.

- CER (Cell Error Ratio) - fraction of cells that are delivered with some bits wrong.
- SECBR (Severely-Errored Cell Block Ratio) - fraction of N-cell blocks of which M or more cells contain an error.
- CMR (Cell Misinsertion Rate) - the number of cells/sec that are delivered to the wrong destination on account of an undetected error in the header.

The third part of the traffic contract tells what constitutes obeying the rules (e.g., if the customer sends one cell too early, does this void the contract? Or, if the carrier fails to meet one of its quality targets for a period of 1 msec, can the customer sue?). So, this part of the contract says how strictly the first two parts will be enforced).

3.3.6. ATM LANs

As it becomes increasingly obvious that replacing the public switched telephone network by an ATM network is going to take a very long time, attention is shifting to the use of ATM technology to connect existing LANs together. In this approach, an ATM network can function either as LAN, connecting individual hosts, or as a bridge, connecting multiple LANs.

The major problem that must be solved is how to provide connectionless LAN service over a connection-oriented ATM network. One possible solution is to introduce a connectionless server into the network. Every host initially sets up a connection to this server, and sends all packets to it for forwarding. While simple, this solution does not use the full bandwidth of the ATM network, and the connectionless server can easily become a bottleneck.

An alternative approach, proposed by ATM Forum is shown in Fig. 5-76. Here every host has a (potential) ATM virtual circuit to every other host. To send a frame, the source host first encapsulates the packet in the payload field of an ATM AAL message and sends it to the destination, the same way frames are shipped over Ethernet and other LANs.

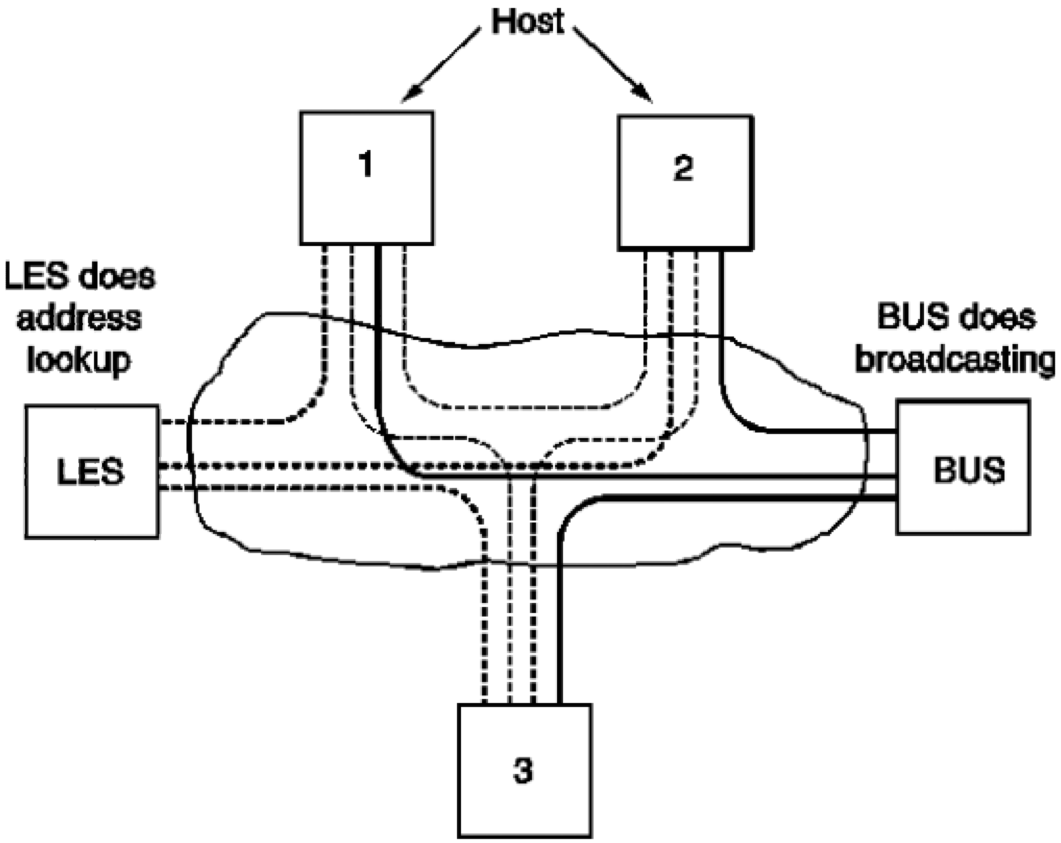


Fig. 5-76. ATM LAN emulation.

The main problem introduced by this scheme is how to tell which IP (or other network layer address) belongs to which virtual circuit. At Ethernet, the problem is solved by ARP protocol using broadcasting, but this does not work with ATM LANs because they do not support broadcasting.

This problem is solved by introducing a new server, the LES (LAN Emulation Server). To look up a network layer address, a host send a packet to the LES, which then looks up the corresponding ATM address and returns it to the machine requesting it.

Some programs use broadcasting or multicasting as an essential part of the application. For these applications, the BUS (Broadcast/Unknown Server) is introduced. It has connection to all hosts and can simulate broadcasting by sending a packet to all of them.

A model similar to this one has been adopted by the IETF (Internet Engineering Task Force) as the official Internet way to use an ATM network for transporting IP packets. In this model the LES server is called ATMARP server. Broadcasting and multicasting are not supported in the IETF proposal. The model is described in RFC 1483 and RFC 1577.

In the IETF method, a set of ATM hosts can be grouped together to form a logical IP subnet (LIS). Each LIS has its own ATMARP server. In effect, a LIS acts like a virtual LAN. Hosts on the same LIS may exchange IP packets directly, but hosts on different ones are required to go through a router. The reason for having LISes is that every host on a LIS must (potentially) have an open virtual circuit to every other host on its LIS. By restricting the number of hosts per LIS, the number of open virtual circuits can be reduced to a manageable number.

Another use of ATM networks is to use them as bridges to connect existing LANs. In this configuration only one machine on each LAN needs an ATM connection.

ATM LAN emulation is an interesting idea, but there are serious questions about its performance and price, and there is certainly heavy competition from existing LANs and bridges, which are well established and highly optimized. Whether ATM LANs and bridges ever replace them remains to see.