CHAPTER -3

Networking and Internetworking

- The networks used in distributed systems are built from a variety of ***transmission media***, including wire, cable, fibre and wireless channels; hardware devices, including routers, switches, bridges, hubs, repeaters and network interfaces; and software components, including protocol stacks, communication handlers and drivers.

- The collection of hardware and software components that provide the communication facilities for a distributed system is referred as ***communication subsystem***.

- In this chapter, the impact of network technologies on the communication subsystem; operating system issues are discussed.

3.1.1 Networking Issues for Distributed Systems

- Performance
- Scalability
- Reliability
- Security
- Mobility
- Quality of Service
- Multicast

1. Performance

- The network performance parameters that are of primary interest for our purposes are those affecting the speed with which individual messages can be transferred between two interconnected computers.
- These are the latency and the point-to-point data transfer rate.
  - Latency is the delay that occurs after a send operation is executed and before data starts to arrive at the destination computer. It can be measured as the time required to transfer an empty message. Here we are considering only network latency, which forms a part of the process-to-process latency.
  - Data transfer rate is the speed at which data can be transferred between two computers in the network once transmission has begun, usually quoted in bits per second.

1. Performance (cont.)

- The time required for a network to transfer a message containing length bits between two computers is:

Message transmission time = latency + length ⁄ data transfer rate

- The above equation is valid for messages whose length does not exceed a maximum that is determined by the underlying network technology. Longer messages have to be segmented and the transmission time is the sum of the times for the segments.
- The transfer rate of a network is determined primarily by its physical characteristics, whereas the latency is determined primarily by software overheads, routing delays and a load-dependent statistical element arising from conflicting demands for access to transmission channels.
- Many of the messages transferred between processes in distributed systems are small in size; latency is therefore often of equal or greater significance than transfer rate in determining performance.

1. Performance (cont.)

- The **total system bandwidth** of a network is a measure of throughput – the total volume of traffic that can be transferred across the network in a given time.

- In many local area network technologies, such as Ethernet, the full transmission capacity of the network is used for every transmission and the system bandwidth is the same as the data transfer rate. But in most wide area networks messages can be transferred on several different channels simultaneously, and the total system bandwidth bears no direct relationship to the transfer rate.

- The performance of networks deteriorates in conditions of overload – when there are too many messages in the network at the same time. The precise effect of overload on the latency, data transfer rate and total system bandwidth of a network depends strongly on the network technology.

1. Performance (cont.)

- Now consider the performance of client-server communication. The time required to transmit a short request message and receive a short reply between nodes on a lightly loaded local network (including system overheads) is about half a millisecond. This should be compared with the sub-microsecond time required to invoke an operation on an application-level object in the local memory.

- **Thus, despite advances in network performance, the time required to access shared resources on a local network remains about a thousand times greater than that required to access resources that are resident in local memory.**

1. Performance (cont.)

- But networks often outperform hard disks; networked access to a local web server or file server with a large in-memory cache of frequently used files can match or outstrip access to files stored on a local hard disk.

- On the Internet, round-trip latencies are in the 5–500 ms range, with means of 20–200 ms depending on distance, so requests transmitted across the Internet are 10–100 times slower than those sent on fast local networks.

- The bulk of this time difference derives from switching delays at routers and contention for network circuits.

2. Scalability

- Computer networks are an indispensable part of the infrastructure of modern societies. The growth since then has been so rapid and diverse that it is difficult to find recent reliable statistics.

- It is realistic to expect it to include several billion nodes and hundreds of millions of active hosts.

- These numbers indicate the future changes in size and load that the Internet must handle. The network technologies on which it is based were not designed to cope with even the Internet's current scale, but they have performed remarkably well. Some substantial changes to the addressing and routing mechanisms are in progress in order to handle the next phase of the Internet's growth.

2. Scalability (cont.)

- For simple client-server applications such as the Web, we would expect future traffic to grow at least in proportion to the number of active users.

- The ability of the Internet's infrastructure to cope with this growth will depend upon the economics of use, in particular charges to users and the patterns of communication that actually occur – for example, their degree of locality.

3. Reliability

- Many applications are able to recover from communication failures and hence do not require guaranteed error-free communication.

- The end-to-end argument further supports the view that the communication subsystem need not provide totally error-free communication; the detection of communication errors and their correction is often best performed by application-level software.

- The reliability of most physical transmission media is very high. When errors occur they are usually due to failures in the software at the sender or receiver (for example, failure by the receiving computer to accept a packet) or buffer overflow rather than errors in the network.

4. Security

- The first level of defence adopted by most organizations is to protect its networks and the computers attached to them with a ***firewall*.** A firewall creates a protection boundary between the organization's intranet and the rest of the Internet.

- The purpose of the firewall is to protect the resources in all of the computers inside the organization from access by external users or processes and to control the use of resources outside the firewall by users inside the organization.

- A firewall runs on a gateway – a computer that stands at the network entry point to an organization's intranet. The firewall receives and filters all of the messages travelling into and out of an organization. It is configured according to the organization's security policy to allow certain incoming and outgoing messages to pass through it and to reject all others.

4. Security (cont.)

- To enable distributed applications to move beyond the restrictions imposed by firewalls there is a need to produce a secure network environment in which a wide range of distributed applications can be deployed, with end-to-end authentication, privacy and security.

- This finer-grained and more flexible form of security can be achieved through the use of cryptographic techniques. It is usually applied at a level above the communication subsystem.

- Exceptions include the need to protect network components such as routers against unauthorized interference with their operation and the need for secure links to mobile devices and other external nodes to enable them to participate in a secure intranet – the ***virtual private network*** (VPN) concept.

5. Mobility

- Mobile devices such as laptop computers and Internet-capable mobile phones are moved frequently between locations and reconnected at convenient network connection points or even used while on the move.

- Wireless networks provide connectivity to such devices, but the addressing and routing schemes of the Internet were developed before the advent of these mobile devices and are not well adapted to their need for intermittent connection to many different subnets.

- The Internet’s mechanisms have been adapted and extended to support mobility, but the expected future growth in the use of mobile devices will demand further development.

6. Quality of Service

- Applications that transmit multimedia data require guaranteed bandwidth and bounded latencies for the communication channels that they use.

- Some applications vary their demands dynamically and specify both a minimum acceptable quality of service and a desired optimum.

7. Multicasting

- Most communication in distributed systems is between pairs of processes, but there often is also a need for one-to-many communication.

- While this can be simulated by *sends* to several destinations, that is more costly than necessary and may not exhibit the fault-tolerance characteristics required by applications.

- For these reasons, many network technologies support the simultaneous transmission of messages to several recipients.

3.2 Types of Network

- The main types of network that are used to support distributed systems: *personal area networks*, *local area networks*, *wide area networks*, *metropolitan area networks* and the wireless variants of them.

- ***Internetworks*** such as the Internet are constructed from networks of all these types.

- Some of the names used to refer to types of networks are confusing because they seem to refer to the physical extent (local area, wide area), but they also identify physical transmission technologies and low-level protocols. These are different for local and wide area networks, although some network technologies, such as ATM (Asynchronous Transfer Mode), are suitable for both local and wide area applications and some wireless networks also support local and metropolitan area transmission.

Types of Network (cont.)

- We refer to networks that are composed of many interconnected networks, integrated to provide a single data communication medium, as internetworks.
- The Internet is the prototypical internetwork; it is composed of millions of local, metropolitan and wide area networks.

Types of Network (cont.)

- Personal area networks (PANs)
- Local area networks (LANs)
- Wide area networks (WANs)
- Metropolitan area networks (MANs)
- Wireless local area networks (WLANs)
- Wireless metropolitan area networks (WMANs)
- Wireless wide area networks (WWANs)
- Internetworks

Figure 3.1 Network performance

	Example	*Range*	*Bandwidth (Mbps)*	*Latency (ms)*
Wired:				
LAN	Ethernet	1–2 kms	10–10,000	1–10
WAN	IP routing	worldwide	0.010–600	100–500
MAN	ATM	2–50 kms	1–600	10
Internetwork	Internet	worldwide	0.5–600	100–500
Wireless:				
WPAN	Bluetooth (IEEE 802.15.1)	10–30m	0.5–2	5–20
WLAN	WiFi (IEEE 802.11)	0.15–1.5 km	11–108	5–20
WMAN	WiMAX (IEEE 802.16)	5–50 km	1.5–20	5–20
WWAN	3G phone	cell: 1--5 km	348–14.4	100–500

Personal area networks (PANs)

- PANs are a subcategory of local networks in which the various digital devices carried by a user are connected by a low-cost, low-energy network.

- Wired PANs are not of much significance because few users wish to be encumbered by a network of wires on their person, but wireless personal area networks (WPANs) are of increasing importance due to the number of personal devices such as mobile phones, tablets, digital cameras, music players and so on that are now carried by many people.

- An example of WPAN is Bluetooth.

Local area networks (LANs)

- LANs carry messages at relatively high speeds between computers connected by a single communication medium, such as twisted copper wire, coaxial cable or optical fibre.

- A *segment* is a section of cable that serves a department or a floor of a building and may have many computers attached. No routing of messages is required within a segment, since the medium provides direct connections between all of the computers connected to it.

- The total system bandwidth is shared between the computers connected to a segment. Larger local networks, such as those that serve a campus or an office building, are composed of many segments interconnected by switches or hubs.

- In local area networks, the total system bandwidth is high and latency is low, except when message traffic is very high.

Local area networks (LANs) (cont.)

- It was originally produced in the early 1970s with a bandwidth of 10 Mbps (million bits per second) and extended to 100 Mbps, 1000 Mbps (1 gigabit per second) and 10 Gbps versions more recently.

- Ethernet technology lacks the latency and bandwidth guarantees needed by many multimedia applications.

- ATM networks were developed to fill this gap, but their cost has inhibited their adoption in local area applications. Instead, high-speed Ethernets have been deployed in a switched mode that overcomes these drawbacks to a significant degree, though not as effectively as ATM.

Wide area networks (WANs)

- WANs carry messages at lower speeds between nodes that are often in different organizations and may be separated by large distances. They may be located in different cities, countries or continents. The communication medium is a set of communication circuits linking a set of dedicated computers called *routers.*

- They manage the communication network and route messages or packets to their destinations. In most networks, the routing operations introduce a delay at each point in the route, so the total latency for the transmission of a message depends on the route that it follows and the traffic loads in the various network segments that it traverses.

Wide area networks (WANs) (cont.)

- In current networks these latencies can be as high as 0.1 to 0.5 seconds. The speed of electronic signals in most media is close to the speed of light, and this sets a lower bound on the transmission latency for long-distance networks. For example, the propagation delay for a signal to travel from Europe to Australia via a terrestrial link is approximately 0.13 seconds and signals via a geostationary satellite between any two points on the Earth’s surface are subject to a delay of approximately 0.20 seconds.
- Bandwidths available across the Internet also vary widely. Speeds of up to 600 Mbps are commonly available, but speeds of 1–10 Mbps are more typically experienced for bulk transfers of data.

Metropolitan area networks (MANs)

- This type of network is based on the high-bandwidth copper and fibre optic cabling recently installed in some towns and cities for the transmission of video, voice and other data over distances of up to 50 kilometres. A variety of technologies have been used to implement the routing of data in MANs, ranging from Ethernet to ATM.

- The DSL (Digital Subscriber Line) and cable modem connections available in many countries are an example. DSL typically uses ATM switches located in telephone exchanges to route digital data onto twisted pairs of copper wire (using high frequency signalling on the existing wiring used for telephone connections) to the subscriber’s home or office at speeds in the range 1–10 Mbps.

- The use of twisted copper wire for DSL subscriber connections limits the range to about 5.5 km from the switch.

Metropolitan area networks (MANs) (cont.)

- Cable modem connections use analogue signalling on cable television networks to achieve speeds of up to 15 Mbps over coaxial cable with greater range than DSL.
- The term DSL actually represents a family of technologies, sometimes referred to as xDSL and including for example ADSL (or Asymmetric Digital Subscriber Line).
- Latest developments include VDSL and VDSL2 (Very High Bit Rate DSL), which are capable of speeds of up to 100 Mbps and designed to support a range of multimedia traffic including High Definition TV (HDTV).

Wireless local area networks (WLANs)

- WLANs are **designed for use in place of wired LANs** to **provide connectivity for mobile devices**, or simply to remove the need for a wired infrastructure to connect computers within homes and office buildings to each other and the Internet.

- They are in widespread use in several variants of the IEEE 802.11 standard (WiFi), offering bandwidths of 10–100 Mbps over ranges up to 1.5 kilometres.

Wireless metropolitan area networks (WMANs)

- The IEEE 802.16 WiMAX standard is targeted at this class of network.

- It aims to provide **an alternative to wired connections to home and office buildings** and to supersede 802.11 WiFi networks in some applications.

Wireless wide area networks (WWANs)

- Most mobile phone networks are based on digital wireless network technologies such as the GSM (Global System for Mobile communication) standard, which is used in most countries of the world.
- **Mobile phone networks are designed to operate over wide areas** (typically entire countries or continents) through the use of cellular radio connections; their data transmission facilities therefore **offer wide area mobile connections to the Internet for portable devices**.

- High speed mobile phone networks are 3G and 4G.

Internetworks

- An internetwork is a communication subsystem in which **several networks are linked together** to provide common data communication facilities **that overlay the technologies and protocols of the individual component networks** and the methods used for their interconnection.
- Internetworks are **needed for the development of extensible, open distributed systems**.
- In internetworks, a variety of local and wide area network technologies can be integrated to provide the networking capacity needed by each group of users.
- The **Internet** is the major instance of internetworking.

3.3 Network Principles

- The basis for all computer networks is the packet-switching technique
- This enables data packets addressed to different destinations to share a single communications link, unlike the circuit-switching technology that underlies conventional telephony. Packets are queued in a buffer and transmitted when the link is available.
- Communication is asynchronous – messages arrive at their destination after a delay that varies depending upon the time that packets take to travel through the network.

- Packet Transmission

- Data Streaming

Packet Transmission

- In most applications of computer networks the requirement is for the transmission of logical units of information, or *messages* – sequences of data items of arbitrary length.

- But **before a message is transmitted it is subdivided into *packets***. The simplest form of packet is a sequence of binary data (an array of bits or bytes) of restricted length, together with addressing information sufficient to identify the source and destination computers. Packets of restricted length are used:

  - so that each computer in the network can allocate sufficient buffer **storage to hold the largest possible incoming packet**;
  - to **avoid the undue delays** that would occur in waiting for communication channels to become free if long messages were transmitted without subdivision.

Data Streaming

- The **transmission** and display of audio and video **in real time is referred to as streaming.**

- It requires much higher bandwidths than most other forms of communication in distributed systems. A video stream requires a bandwidth of about 1.5 Mbps if the data is compressed, or 120 Mbps if uncompressed.
- UDP internet packets are generally used to hold the video frames, but because the flow is continuous as opposed to the intermittent traffic generated by typical client-server interactions, the packets are handled somewhat differently.
- The *play time* of a multimedia element such as a video frame is the time at which it must be displayed (for a video element) or converted to sound (for a sound sample).

Data Streaming (cont.)

- For example, in a stream of video frames with a frame rate of 24 frames per second, frame *N* has a play time that is *N*/24 seconds after the stream's start time.
- Elements that arrive at their destination later than their play time are no longer useful and will be dropped by the receiving process.
- The **timely delivery** of audio and video streams depends upon the availability of connections with adequate quality of service – **bandwidth, latency and reliability must all be considered**. Ideally, adequate quality of service should be guaranteed. In general the Internet does not offer that capability, and the quality of real-time video streams sometimes reflects that, but in proprietary intranets such as those operated by media companies, guarantees are sometimes achieved.

Data Streaming (cont.)

- What is required is the ability to establish a channel from the source to the destination of a multimedia stream, with a predefined route through the network, a reserved set of resources at each node through which it will travel and buffering where appropriate to smooth any irregularities in the flow of data through the channel. Data can then be passed through the channel from sender to receiver at the required rate.

- ATM networks are specifically designed to provide high bandwidth and low latencies and to support quality of service by the reservation of network resources. IPv6, the new network protocol for the Internet includes features that enable each of the IP packets in a real-time stream to be identified and treated separately from other data at the network level.

Data Streaming (cont.)

- **Communication subsystems that provide quality of service guarantees require facilities for the pre-allocation of network resources** and the enforcement of the allocations.
- The **Resource Reservation Protocol (RSVP)** [1993] enables applications to negotiate the pre-allocation of bandwidth for real-time data streams.
- The **Real Time Transport Protocol (RTP)** [1996] is an application-level data transfer protocol that includes details of the play time and other timing requirements in each packet. The availability of effective implementations of these protocols in the general Internet will depend upon substantial changes to the transport and network layers.

3.3 Switching Schemes

- A network consists of a set of nodes connected together by circuits. To transmit information between two arbitrary nodes, a switching system is required.

- Here we define the four types of switching that are used in computer networking:
  1. Broadcasting
  2. Circuit Switching
  3. Packet Switching
  4. Frame Relay

1. Broadcasting

- Broadcasting is a transmission technique that involves no switching.
- Everything is transmitted to every node, and it is up to potential receivers to notice transmissions addressed to them.
- Some LAN technologies, including Ethernet, are based on broadcasting.
- Wireless networking is necessarily based on broadcasting, but in the absence of fixed circuits the broadcasts are arranged to reach nodes grouped in *cells*.

2. Circuit Switching

- At one time telephone networks were the only telecommunication networks. Their operation was simple to understand: when a caller dialled a number, the pair of wires from her phone to the local exchange was connected by an automatic switch at the exchange to the pair of wires connected to the other party's phone.
- For a long-distance call the process was similar but the connection would be switched through a number of intervening exchanges to its destination. This system is sometimes referred to as the *plain old telephone system*, or POTS. It is a typical *circuit-switching network*.

3. Packet Switching

- The advent of computers and digital technology brought many new possibilities for telecommunication. At the most basic level, it brought processing and storage. These made it possible to construct a different kind of communication network called a ***store-and-forward network.***

- Instead of making and breaking connections to build circuits, a store-and-forward network just forwards packets from their source to their destination. There is a computer at each switching node (that is, wherever several circuits need to be interconnected). Each packet arriving at a node is first stored in memory at the node and then processed by a program that transmits it on an outgoing circuit, which transfers the packet to another node that is closer to its ultimate destination.

4. Frame Relay

- it takes anything from a few tens of microseconds to a few milliseconds to switch a packet through each network node in a store-and-forward network.
- This switching delay depends on the packet size, hardware speed and quantity of other traffic, but its lower bound is determined by the network bandwidth, since the entire packet must be received before it can be forwarded to another node.
- Much of the Internet is based on store-and-forward switching, and as we have already seen, even short Internet packets typically take up to 200 milliseconds to reach their destinations.
- Delays of this magnitude are too long for real-time applications such as telephony and video conferencing.

4. Frame Relay (cont.)

- The *frame relay* switching method brings some of the advantages of circuit switching to packet-switching networks. They overcome the delay problems by switching small packets (called *frames*) on the fly.
- The switching nodes (which are usually special-purpose parallel digital processors) route frames based on the examination of their first few bits; frames as a whole are not stored at nodes but pass through them as short streams of bits.
- ATM networks are a prime example; high-speed ATM networks can transmit packets across networks consisting of many nodes in a few tens of microseconds.

Packet Delivery

- There are two approaches to the delivery of packets by the network layer:
  1. Datagram packet delivery
  2. Virtual circuit packet delivery

1. Datagram packet delivery:

- The term ‘datagram’ refers to the similarity of this delivery mode to the way in which letters and telegrams are delivered. The essential feature of datagram networks is that the delivery of each packet is a ‘one-shot’ process; no setup is required, and once the packet is delivered the network retains no information about it. In a datagram network a sequence of packets transmitted by a single host to a single destination may follow different routes (if, for example, the network is capable of adaptation to handle failures or to mitigate the effects of localized congestion), and when this occurs they may arrive out of sequence.

1. Datagram packet delivery (cont.)

- Every datagram packet contains the full network address of the source and destination hosts; the latter is an essential parameter for the routing process, which we describe in the next section.
- Datagram delivery is the concept on which packet networks were originally based, and it can be found in most of the computer networks in use today.
- The Internet’s network layer (IP), Ethernet and most wired and wireless local network technologies are based on datagram delivery.

2. Virtual circuit packet delivery

- Some network-level services implement packet transmission in a manner that is analogous to a telephone network. A virtual circuit must be set up before packets can pass from a source host A to destination host B. The establishment of a virtual circuit involves the identification of a route from the source to the destination, possibly passing through several intermediate nodes. At each node along the route a table entry is made, indicating which link should be used for the next stage of the route.
- Once a virtual circuit has been set up, it can be used to transmit any number of packets. Each network-layer packet contains only a virtual circuit number in place of the source and destination addresses. The addresses are not needed, because packets are routed at intermediate nodes by reference to the virtual circuit number. When a packet reaches its destination the source can be determined from the virtual circuit number.

2. Virtual circuit packet delivery (cont.)

- The analogy with telephone networks should not be taken too literally. In the POTS a telephone call results in the establishment of a physical circuit from the caller to the callee, and the voice links from which it is constructed are reserved for their exclusive use.
- In virtual circuit packet delivery the circuits are represented only by table entries in routing nodes, and the links along which the packets are routed are used only for the time taken to transmit a packet; they are free for other uses for the rest of the time. A single link may therefore be employed in many separate virtual circuits.
- The most important virtual circuit network technology in current use is ATM. it benefits from lower latencies for the transmission of individual packets; this is a direct result of its use of virtual circuits. The requirement for a setup phase does, however, result in a short delay before any packets can be sent to a new destination.
- The distinction between datagram and virtual circuit packet delivery in the network layer should not be confused with a similarly named pair of mechanisms in the transport layer: connectionless and connection-oriented transmission.

There are two approaches to the delivery of packets by the network layer:

Datagram packet delivery: The term ‘datagram’ refers to the similarity of this delivery mode to the way in which letters and telegrams are delivered. The essential feature of datagram networks is that the delivery of each packet is a ‘one-shot’ process; no setup is required, and once the packet is delivered the network retains no information about it. In a datagram network a sequence of packets transmitted by a single host to a single destination may follow different routes

Virtual circuit packet delivery: A virtual circuit must be set up before packets can pass from a source host A to destination host B. The establishment of a virtual circuit involves the identification of a route from the source to the destination, possibly passing through several intermediate nodes. At each node along the route a table entry is made, indicating which link should be used for the next stage of the route.

The delivery of packets to their destinations in a network such as the one shown in Figure 3.7 is the collective responsibility of the routers located at connection points.

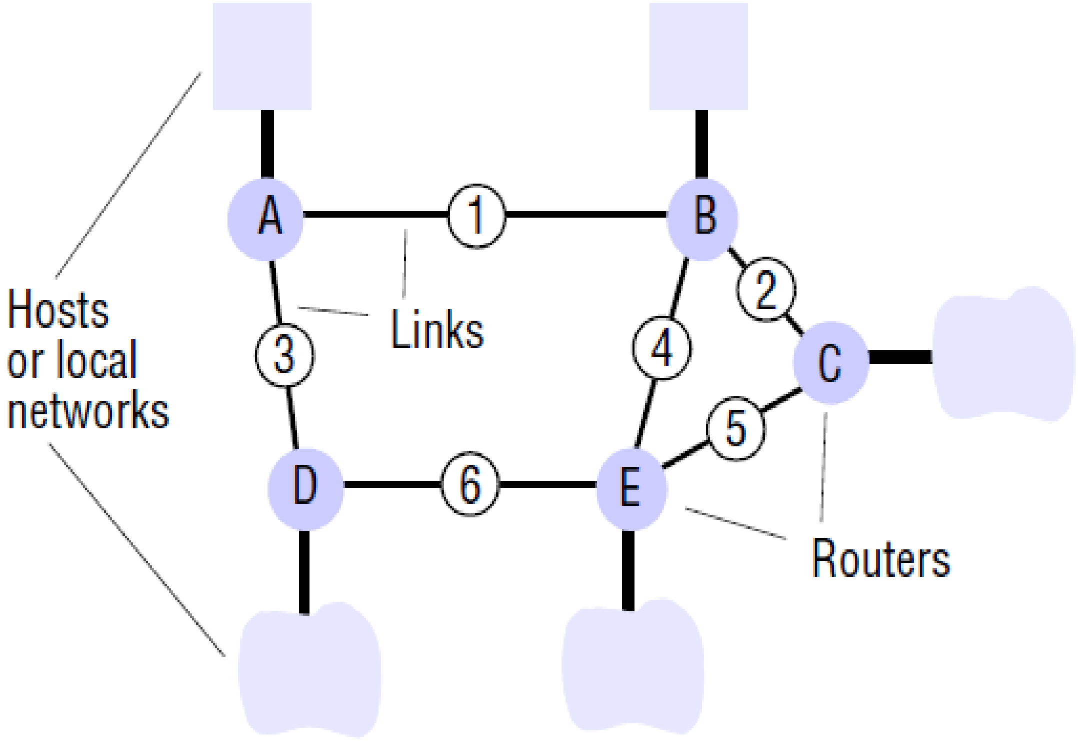


Fig 3.7 Routing in a wide area network

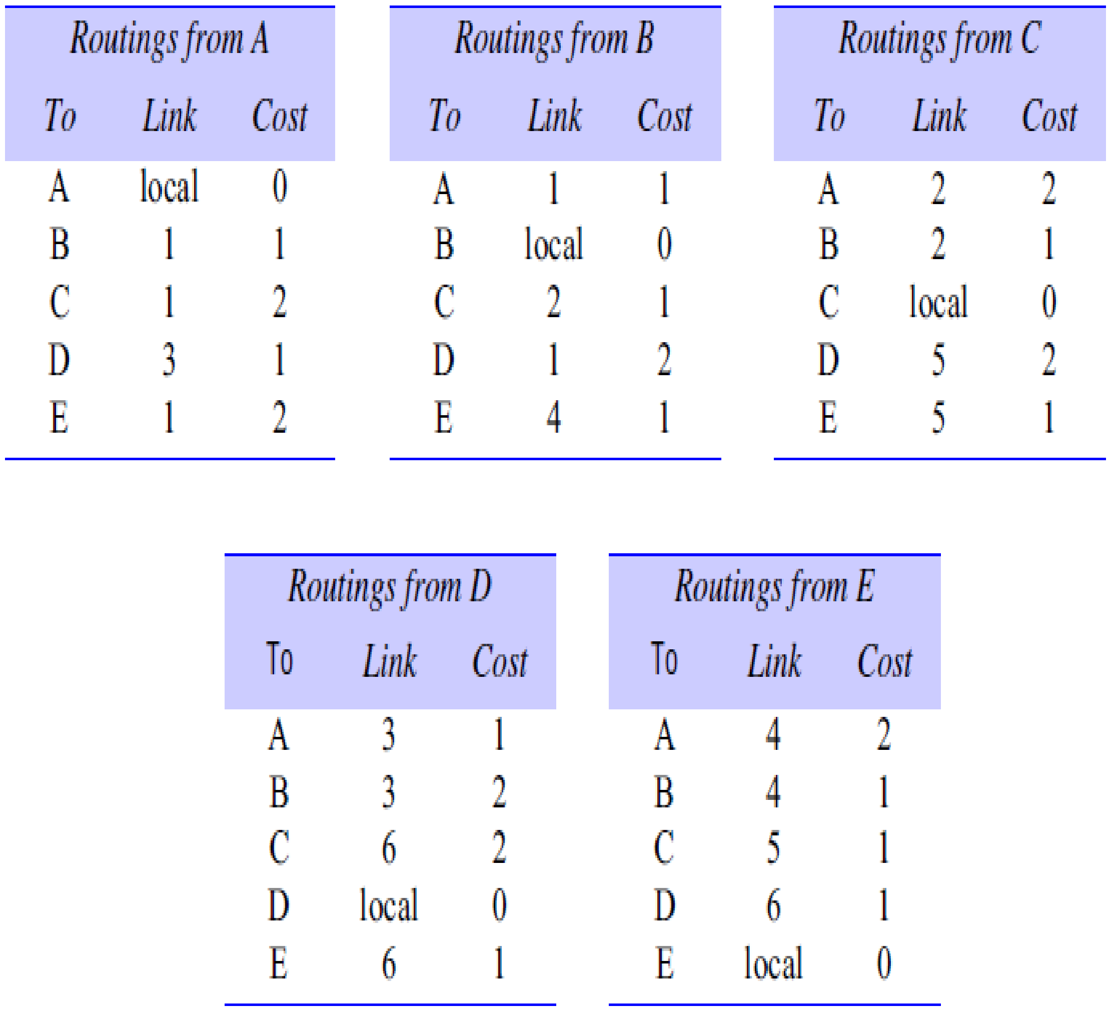
A routing algorithm has two parts:

1. It must make decisions that determine the route taken by each packet as it travels through the network.

2. It must dynamically update its knowledge of the network based on traffic monitoring and the detection of configuration changes or failures. This activity is less time-critical; slower and more computation-intensive techniques can be used

Both of these activities are distributed throughout the network. The routing decisions are made on a hop-by-hop basis, using locally held information to determine the next hop to be taken by each incoming packet.

Figure 3.8 Routing tables for the network in Figure 3.7



Routings from A		
To	*Link*	*Cost*
A	local	0
B	1	1
C	1	2
D	3	1
E	1	2

Routings from B		
To	*Link*	*Cost*
A	1	1
B	local	0
C	2	1
D	1	2
E	4	1

Routings from C		
To	*Link*	*Cost*
A	2	2
B	2	1
C	local	0
D	5	2
E	5	1

Routings from D		
To	*Link*	*Cost*
A	3	1
B	3	2
C	6	2
D	local	0
E	6	1

Routings from E		
To	*Link*	*Cost*
A	4	2
B	4	1
C	5	1
D	6	1
E	local	0

A simple routing algorithm

- distance vector’ algorithm
- *link-state* algorithm
- Bellman’s shortest path algorithm

A router exchanges information about the network with its neighbouring nodes by sending a summary of its routing table using a *router information protocol* (RIP). The RIP actions performed at a router are described
informally as follows:

1. *Periodically, and whenever the local routing table changes*

2. *When a table is received from a neighbouring router*, if the received table shows a route to a new destination, or a better (lower-cost) route to an existing destination, update the local table with the new route.

Figure 3.9 Pseudo-code for RIP routing algorithm

```
Send: Each t seconds or when Tl changes, send Tl on each non-faulty outgoing link.

Receive: Whenever a routing table Tr is received on link n:
    for all rows Rr in Tr {
        if ($Rr.link \neq n$) {
            $Rr.cost = Rr.cost + 1$;
            $Rr.link = n$;
            if (Rr.destination is not in Tl) add Rr to Tl;    // add new destination to Tl
            else for all rows Rl in Tl {
                if ($Rr.destination = Rl.destination$ and
                    ($Rr.cost < Rl.cost$ or $Rl.link = n$)) $Rl = Rr$;
                    // $Rr.cost < Rl.cost$ : remote node has better route
                    // $Rl.link = n$ : remote node is more authoritative
            }
        }
    }
```

To deal with faults, each router monitors its links and acts as follows:

When a faulty link n *is detected*, set *cost* to ⍰ for all entries in the local table that refer to the faulty link and perform the *Send* action.

Thus the information that the link is broken is represented by an infinite value for the cost to the relevant destinations. When this information is propagated to neighbouring nodes it will be processed according to the *Receive* action and then propagated further until a node is reached that has a working route to the relevant destinations, if one exists.

The node that still has a working route will eventually propagate its table, and the working route will replace the faulty one at all nodes.

The vector-distance algorithm can be improved in various ways: costs (also known as the *metric*) can be based on the actual bandwidths of the links and the algorithm can be modified to increase its speed of convergence and to avoid some undesirable intermediate states, such as loops, that may occur before convergence is achieved.

3.3.4 Protocols

- The term *protocol* is used to refer to a well-known set of rules and formats to be used for communication between processes in order to perform a given task. The definition of a protocol has two important parts to it:
  - a specification of the sequence of messages that must be exchanged;
  - a specification of the format of the data in the messages.
- The existence of well-known protocols enables the separate software components of distributed systems to be developed independently and implemented in different programming languages on computers that may have different order codes and data representations.

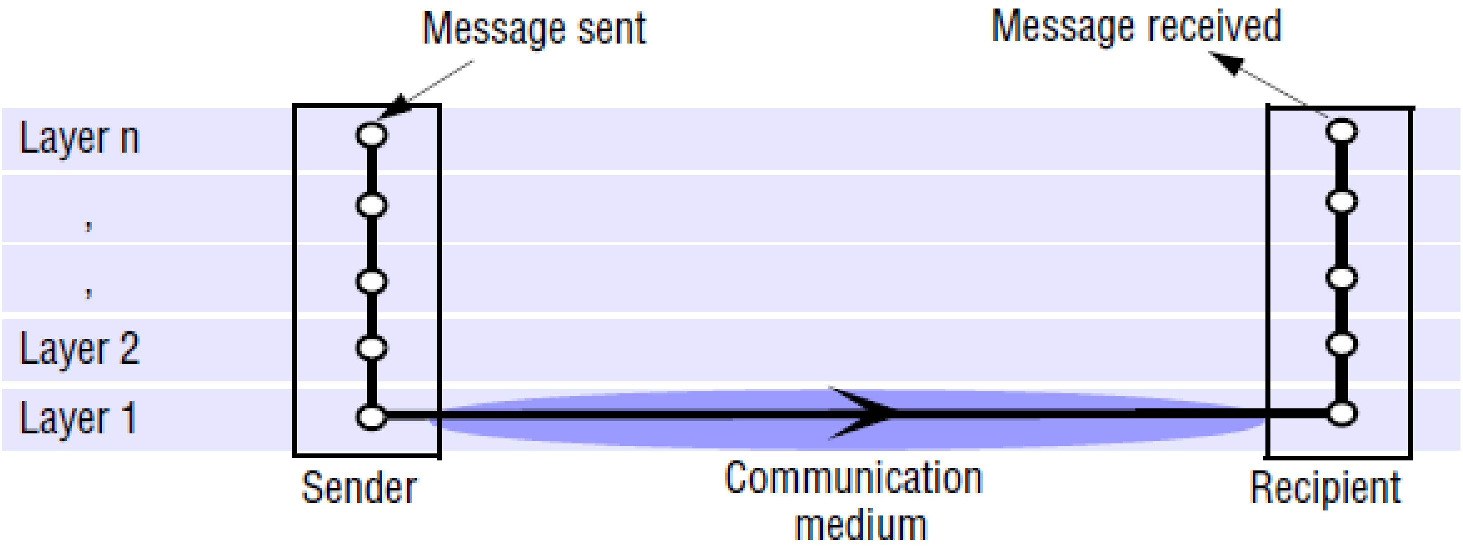
Protocols (cont.)

- A protocol is implemented by a pair of software modules located in the sending and receiving computers.
- For example, a *transport protocol* transmits messages of any length from a sending process to a receiving process. A process wishing to transmit a message to another process issues a call to a transport protocol module, passing it a message in the specified format. The transport software then concerns itself with the transmission of the message to its destination, subdividing it into packets of some specified size and format that can be transmitted to the destination via the *network protocol* – another, lower-level protocol.
- The corresponding transport protocol module in the receiving computer receives the packet via the network-level protocol module and performs inverse transformations to regenerate the message before passing it to a receiving process.

Protocol Layers

- Network software is arranged in a hierarchy of layers. Each layer presents an interface to the layers above it that extends the properties of the underlying communication system. A layer is represented by a module in every computer connected to the network.

Conceptual layering of protocol software



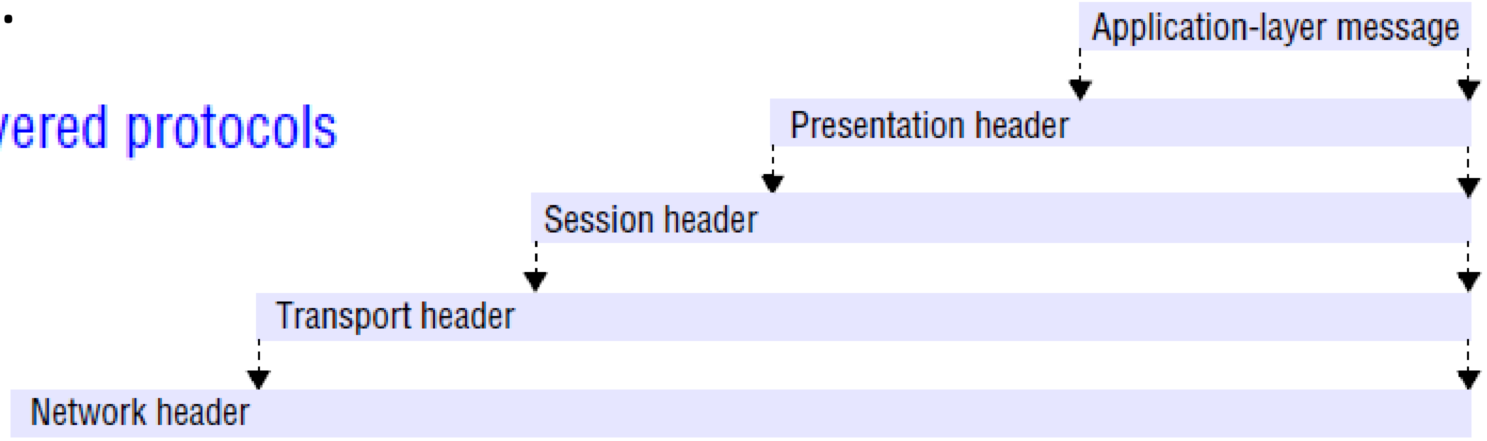
Protocol Layers (cont.)

- The previous figure illustrates the structure and the flow of data when a message is transmitted using a layered protocol. Each module appears to communicate directly with a module at the same level in another computer in the network, but in reality data is not transmitted directly between the protocol modules at each level. Instead, each layer of network software communicates by local procedure calls with the layers above and below it.
- On the sending side, each layer (except the topmost, or *application layer*) accepts items of data in a specified format from the layer above it and applies transformations to encapsulate the data in the format specified for that layer before passing it to the layer below for further processing.

Protocol Layers (cont.)

- Figure illustrates this process as it applies to the top four layers of the OSI protocol suite (discussed in the next subsection). The figure shows the packet headers that hold most network-related data items, but for clarity it omits the trailers that are present in some types of packet; it also assumes that the application-layer message to be transmitted is shorter than the underlying network's maximum packet size. If not, it would have to be encapsulated in several network-layer packets. On the receiving side, the converse transformations are applied to data items received from the layer below before they are passed to the layer above.

Encapsulation as it is applied in layered protocols



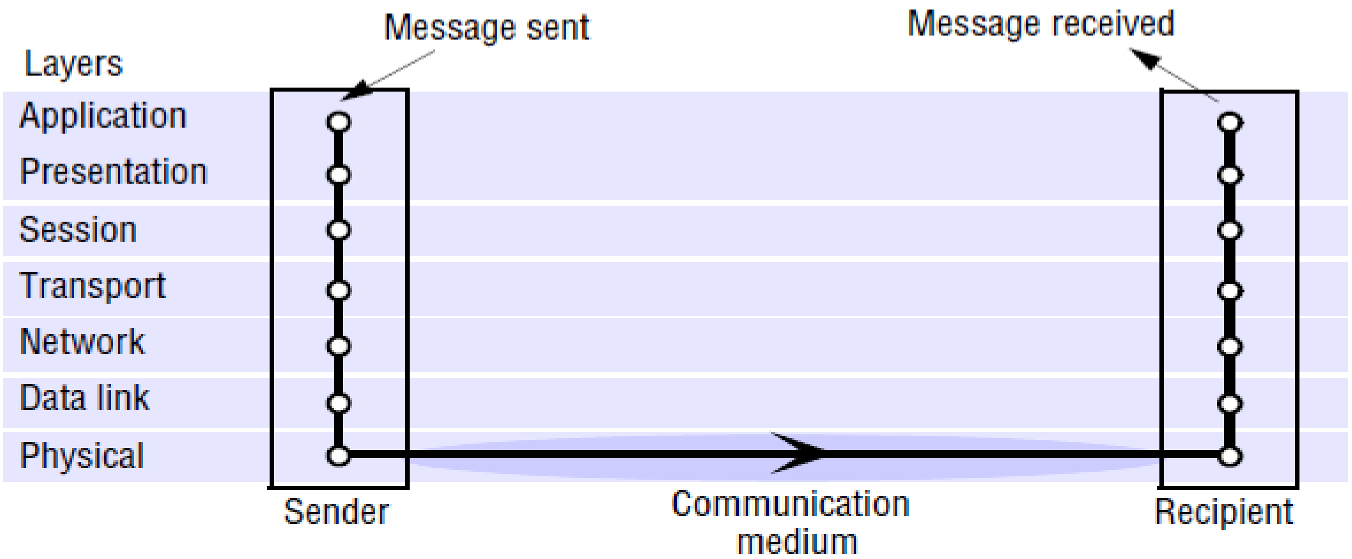
Protocol Layers (cont.)

- The protocol type of the layer above is included in the header of each layer, to enable the protocol stack at the receiver to select the correct software components to unpack the packets.
- Thus each layer provides a service to the layer above it and extends the service provided by the layer below it. At the bottom is a *physical layer*. This is implemented by a communication medium (copper or fibre optic cables, satellite communication channels or radio transmission) and by analogue signalling circuits that place signals on the communication medium at the sending node and sense them at the receiving node.
- At receiving nodes data items are received and passed upwards through the hierarchy of software modules, being transformed at each stage until they are in a form that can be passed to the intended recipient process.

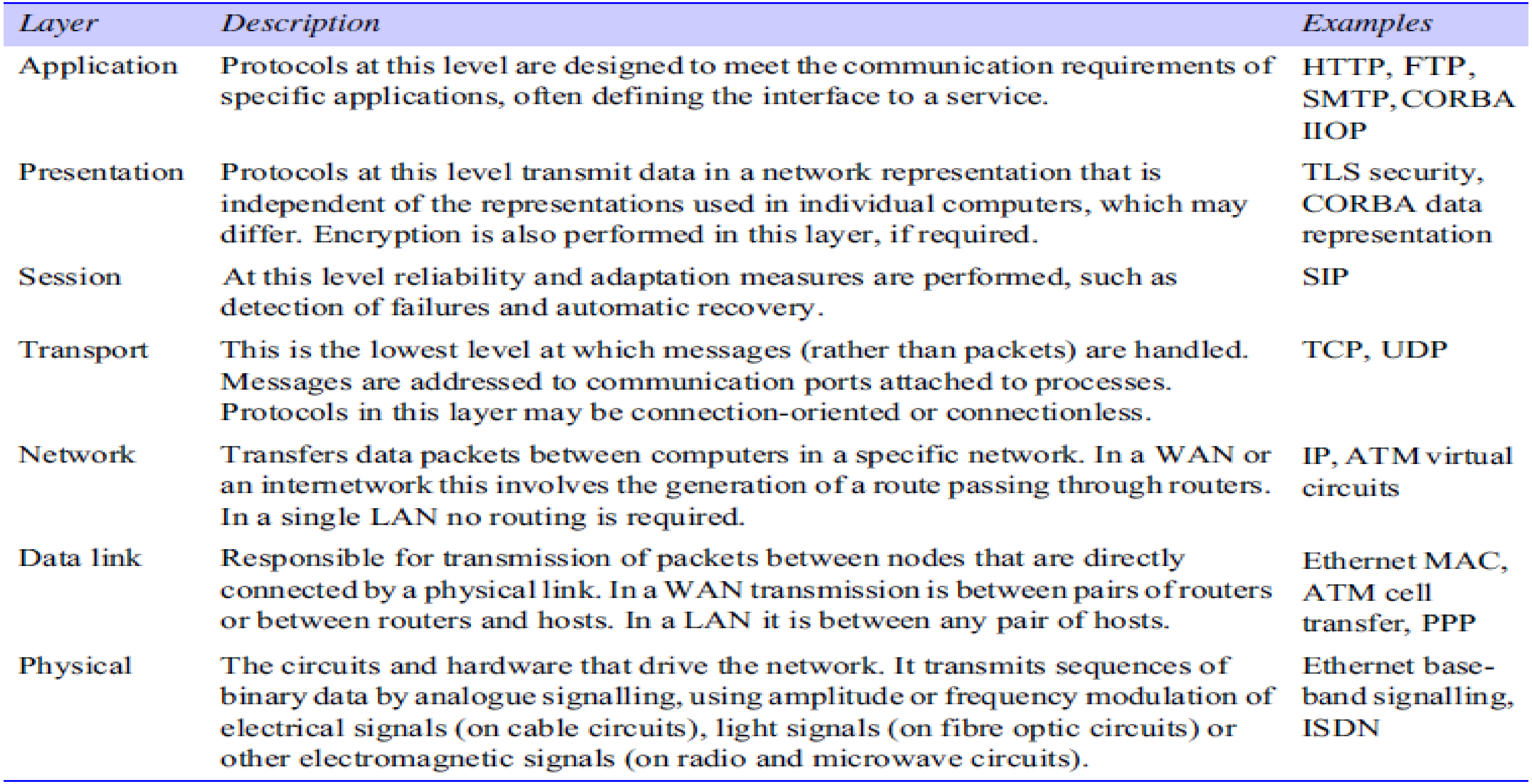
Protocol Suits

- A complete set of protocol layers is referred to as a *protocol suite* or a *protocol stack*, reflecting the layered structure. Figure shows a protocol stack that conforms to the seven-layer Reference Model for *Open Systems Interconnection* (OSI) adopted by the International Organization for Standardization (ISO) [ISO 1992]. The OSI Reference Model was adopted in order to encourage the development of protocol standards that would meet the requirements of open systems.

Protocol layers in the ISO Open Systems Interconnection (OSI) protocol model



OSI protocol summary



Layer	*Description*	*Examples*
Application	Protocols at this level are designed to meet the communication requirements of specific applications, often defining the interface to a service.	HTTP, FTP, SMTP, CORBA IIOP
Presentation	Protocols at this level transmit data in a network representation that is independent of the representations used in individual computers, which may differ. Encryption is also performed in this layer, if required.	TLS security, CORBA data representation
Session	At this level reliability and adaptation measures are performed, such as detection of failures and automatic recovery.	SIP
Transport	This is the lowest level at which messages (rather than packets) are handled. Messages are addressed to communication ports attached to processes. Protocols in this layer may be connection-oriented or connectionless.	TCP, UDP
Network	Transfers data packets between computers in a specific network. In a WAN or an internetwork this involves the generation of a route passing through routers. In a single LAN no routing is required.	IP, ATM virtual circuits
Data link	Responsible for transmission of packets between nodes that are directly connected by a physical link. In a WAN transmission is between pairs of routers or between routers and hosts. In a LAN it is between any pair of hosts.	Ethernet MAC, ATM cell transfer, PPP
Physical	The circuits and hardware that drive the network. It transmits sequences of binary data by analogue signalling, using amplitude or frequency modulation of electrical signals (on cable circuits), light signals (on fibre optic circuits) or other electromagnetic signals (on radio and microwave circuits).	Ethernet base-band signalling, ISDN

OSI Reference Model (cont.)

- It is a framework for the definition of protocols and not a definition for a specific suite of protocols. Protocol suites that conform to the OSI model must include at least one specific protocol at each of the seven levels that the model defines.
- Protocol layering brings substantial benefits in simplifying and generalizing the software interfaces for access to the communication services of networks, but it also carries significant performance costs.
- The transmission of an application-level message via a protocol stack with *N* layers typically involves *N* transfers of control to the relevant layer of software in the protocol suite, at least one of which is an operating system entry, and taking *N* copies of the data as a part of the encapsulation mechanism. All of these overheads result in data transfer rates between application processes that are much lower than the available network bandwidth.

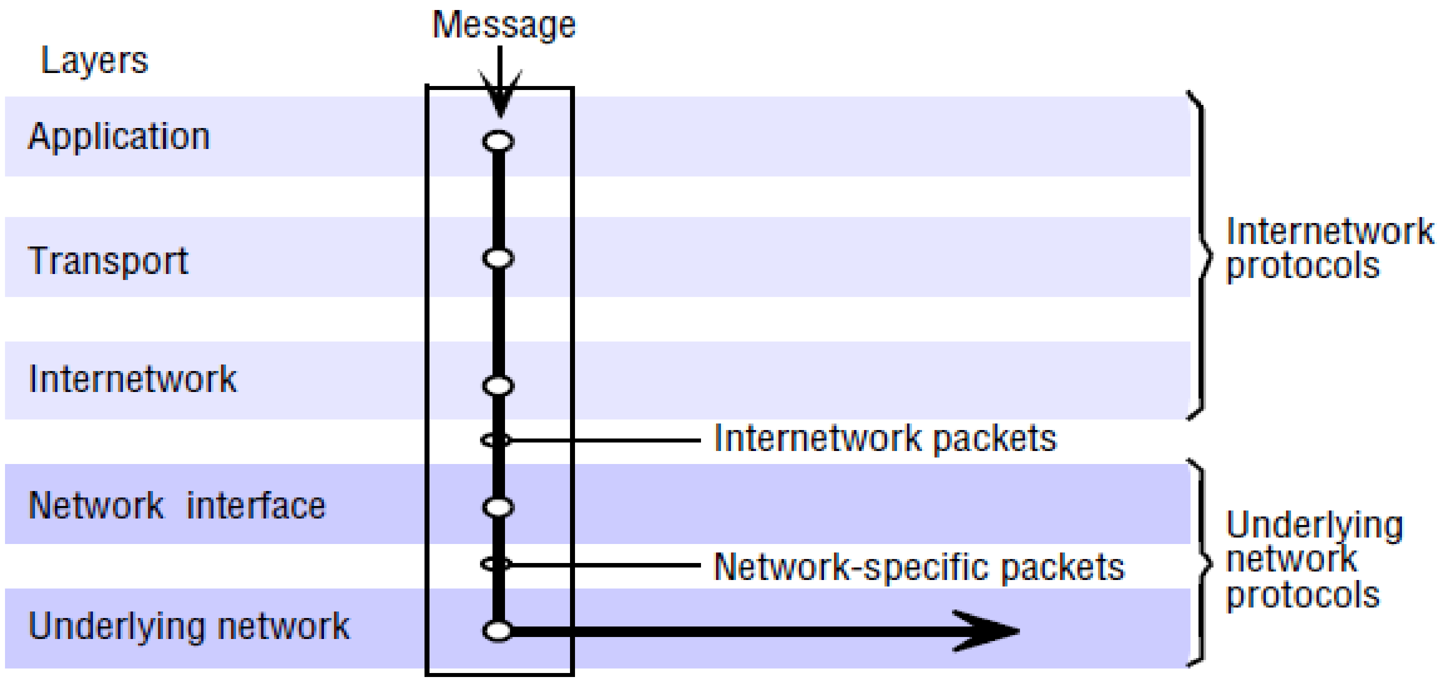
OSI Reference Model (cont.)

- The implementation of the Internet does not follow the OSI model in two respects:
  - First, the application, presentation and session layers are not clearly distinguished in the Internet protocol stack. Instead, the application and presentation layers are implemented either as a single middleware layer or separately within each application. Thus CORBA implements inter-object invocations and data representations in a middleware library that is included in each application process. Web browsers and other applications that require secure channels employ the Secure Sockets Layer as a procedure library in a similar manner.
  - Second, the session layer is integrated with the transport layer. Internetwork protocol suites include an application layer, a transport layer and an *internetwork layer*. The internetwork layer is a ‘virtual’ network layer that is responsible for transmitting internetwork packets to a destination computer. An *internetwork packet* is the unit of data transmitted over an internetwork.

OSI Reference Model (cont.)

- Internetwork protocols are overlaid on underlying networks as illustrated in Figure. The *network interface* layer accepts internetwork packets and converts them into packets suitable for transmission by the network layers of each underlying network.

Internetwork layers



Packet Assembly

- The task of dividing messages into packets before transmission and reassembling them at the receiving computer is usually performed in the transport layer.
- The network-layer protocol packets consist of a *header* and a *data field*. In most network technologies, the data field is variable in length, with the maximum length called the *maximum transfer unit* (MTU).
- If the length of a message exceeds the MTU of the underlying network layer, it must be fragmented into chunks of the appropriate size, with sequence numbers for use on reassembly, and transmitted in multiple packets.
- For example, the MTU for Ethernets is 1500 bytes – no more than that quantity of data can be transmitted in a single Ethernet packet.

Packet Assembly (cont.)

- Although the IP protocol stands in the position of a network-layer protocol in the Internet suite of protocols, its MTU is unusually large at 64 kbytes (8 kbytes is often used in practice because some nodes are unable to handle such large packets).
- Whichever MTU value is adopted for IP packets, packets larger than the Ethernet MTU can arise and they must be fragmented for transmission over Ethernets.

Ports

- The transport layer’s task is to provide a network-independent message transport service between pairs of network *ports*. Ports are software-defined destination points at a host computer. They are attached to processes, enabling data transmission to be addressed to a specific process at a destination node. Next, we discuss the addressing of ports as they are implemented in the Internet and most other networks.

Addressing

- The transport layer is responsible for delivering messages to destinations with *transport addresses* that are composed of the *network address* of a host computer and a *port number*. A network address is a numeric identifier that uniquely identifies a host computer and enables it to be located by nodes that are responsible for routing data to it. In the Internet every host computer is assigned an IP number, which identifies it and the subnet to which it is connected, enabling data to be routed to it from any other node (as described in the following sections). In Ethernets there are no routing nodes; each host is responsible for recognizing and picking up packets addressed to it.

Addressing (cont.)

- Well-known Internet services such as HTTP and FTP have been allocated *contact port numbers* and these are registered with a central authority (the Internet Assigned Numbers Authority (IANA) [www.iana.org I]).
- To access a service at a given host, a request is sent to the relevant port at the host. Some services, such as FTP (contact port: 21), then allocate a new port (with a private number) and send the number of the new port to the client.
- The client uses the new port for the remainder of a transaction or a session. Other services, such as HTTP (contact port: 80), transact all of their business through the contact port.

Addressing (cont.)

- Port numbers below 1023 are defined as *well-known ports* whose use is restricted to privileged processes in most operating systems. The ports between 1024 and 49151 are *registered ports* for which IANA holds service descriptions, and the remaining ports up to 65535 are available for private purposes. In practice, all of the ports above 1023 can be used for private purposes, but computers using them for private purposes cannot simultaneously access the corresponding registered services.
- A fixed port number allocation does not provide an adequate basis for the development of distributed systems which often include a multiplicity of servers including dynamically allocated ones. Solutions to this problem involve the dynamic allocation of ports to services and the provision of binding mechanisms to enable clients to locate services and their ports using symbolic names

3.3.6 Congestion control

- The capacity of a network is limited by the performance of its communication links and switching nodes. When the load at any particular link or node approaches its capacity, queues will build up at hosts trying to send packets and at intermediate nodes holding packets whose onward transmission is blocked by other traffic. If the load continues at the same high level, the queues will continue to grow until they reach the limit of available buffer space.
- Once this state is reached at a node, the node has no option but to drop further incoming packets.
- As a rule of thumb, when the load on a network exceeds 80% of its capacity, the total throughput tends to drop as a result of packet losses unless usage of heavily loaded links is controlled.
- In general, congestion control is achieved by informing nodes along a route that congestion has occurred and that their rate of packet transmission should therefore be reduced.
- All datagram-based network layers, including IP and Ethernets, rely on the endto-end control of traffic. That is, the sending node must reduce the rate at which it transmits packets based only on information that it receives from the receiver.
- Congestion information may be supplied to the sending node by explicit transmission of special messages (called *choke packets*) requesting a reduction in transmission rate, by the implementation of a specific transmission control protocol

3.3.7 Internetworking

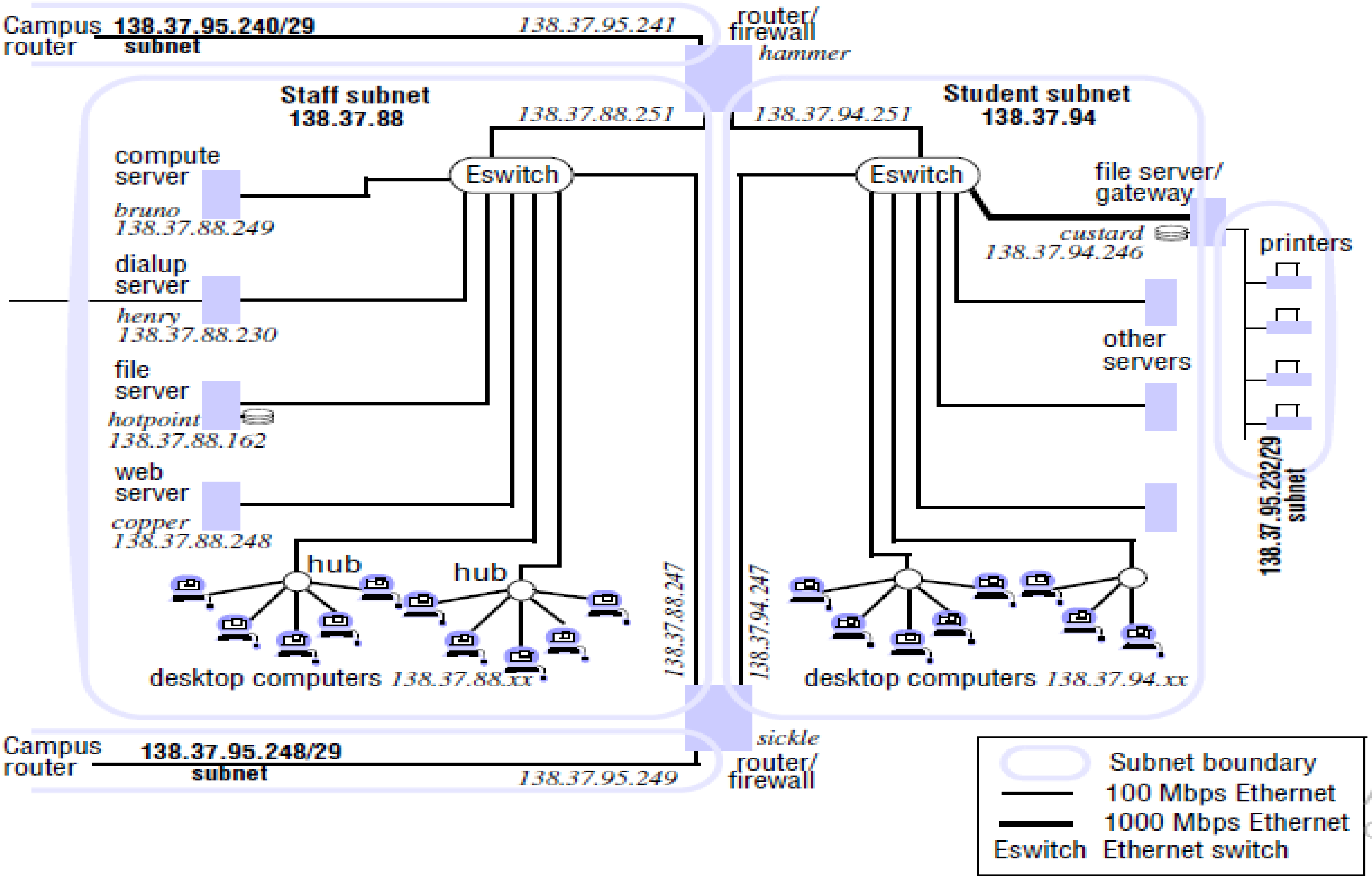
To build an integrated network (an *internetwork*) we must integrate many subnets, each of which is based on one of these network technologies. To make this possible, the following are needed:

1. a unified internetwork addressing scheme that enables packets to be addressed to any host connected to any subnet;

2. a protocol defining the format of internetwork packets and giving rules according to which they are handled;

3. interconnecting components that route packets to their destinations in terms of internetwork addresses, transmitting the packets using subnets with a variety of network technologies.

Figure 3.10 Simplified view of part of a university campus network



The routers (hostnames: *hammer* and *sickle*) are, in fact, general-purpose computers that also fulfil other purposes. One of those purposes is to serve as firewalls; the role of a firewall is closely linked with the routing function.

The 138.37.95.232/29 subnet is not connected to the rest of the network at the IP level. Only the file server *custard* can access it to provide a printing service on the attached printers via a server process that monitors and controls the use of the printers.

All of the links in Figure 3.10 are Ethernets. The bandwidth of most of them is 100 Mbps, but one is 1000 Mbps because it carries a large volume of traffic between a large number of computers used by students and *custard,* the file server that holds all of their files.

There are two Ethernet switches and several Ethernet hubs in the portion of the network illustrated. Both types of component are transparent to IP packets. An Ethernet hub is simply a means of connecting together several segments of Ethernet cable, all of which form a single Ethernet at the network protocol level.

The routers are responsible for forwarding the internetwork packets that arrive on any connection to the correct outgoing connection, as explained above. They maintain routing tables for that purpose.

Bridges link networks of different types. Some bridges link several networks, and these are referred to as bridge/routers because they also perform routing functions.

Hubs are simply a convenient means of connecting hosts and extending segments of Ethernet and other broadcast local network technologies. They have a number of sockets (typically 4–64), to each of which a host computer can be connected

Switches • Switches perform a similar function to routers, but for local networks (normally Ethernets) only. That is, they interconnect several separate Ethernets, routing the incoming packets to the appropriate outgoing network.

The advantage of switches over hubs is that they separate the incoming traffic and transmit it only on the relevant outgoing network, reducing congestion on the other networks to which they are connected.

Tunnelling • Bridges and routers transmit internetwork packets over a variety of underlying networks by translating between their network-layer protocols and an internetwork protocol, but there is one situation in which the underlying network protocol can be hidden from the layers above it without the use of an internetwork protocol. A pair of nodes connected to separate networks of the same type can communicate through another type of network by constructing a protocol 'tunnel'. A protocol tunnel is a software layer that transmits packets through an alien network environment.

3.4.1 IP addressing

Figure 3.15 Internet address structure, showing field sizes in bits

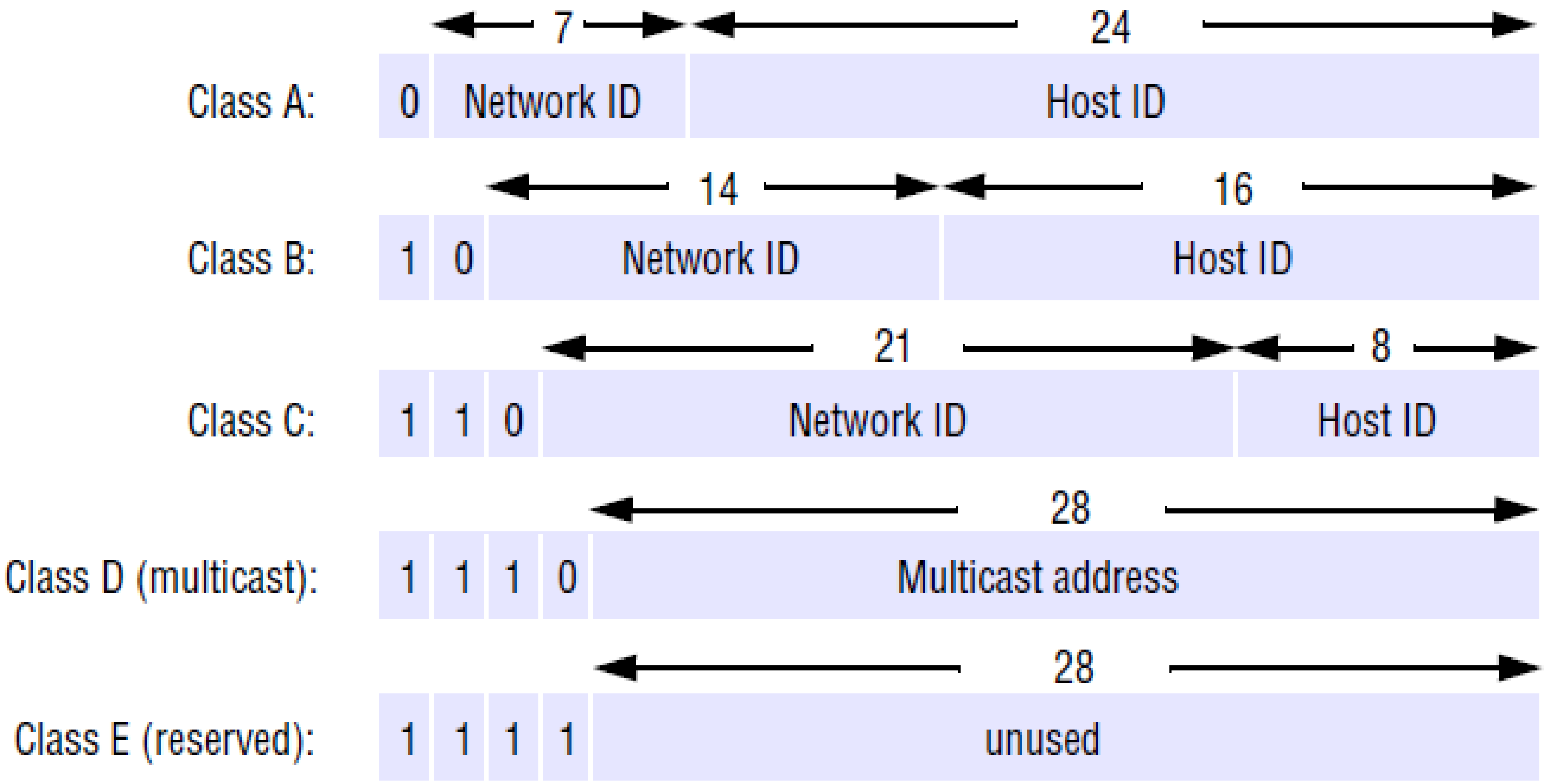
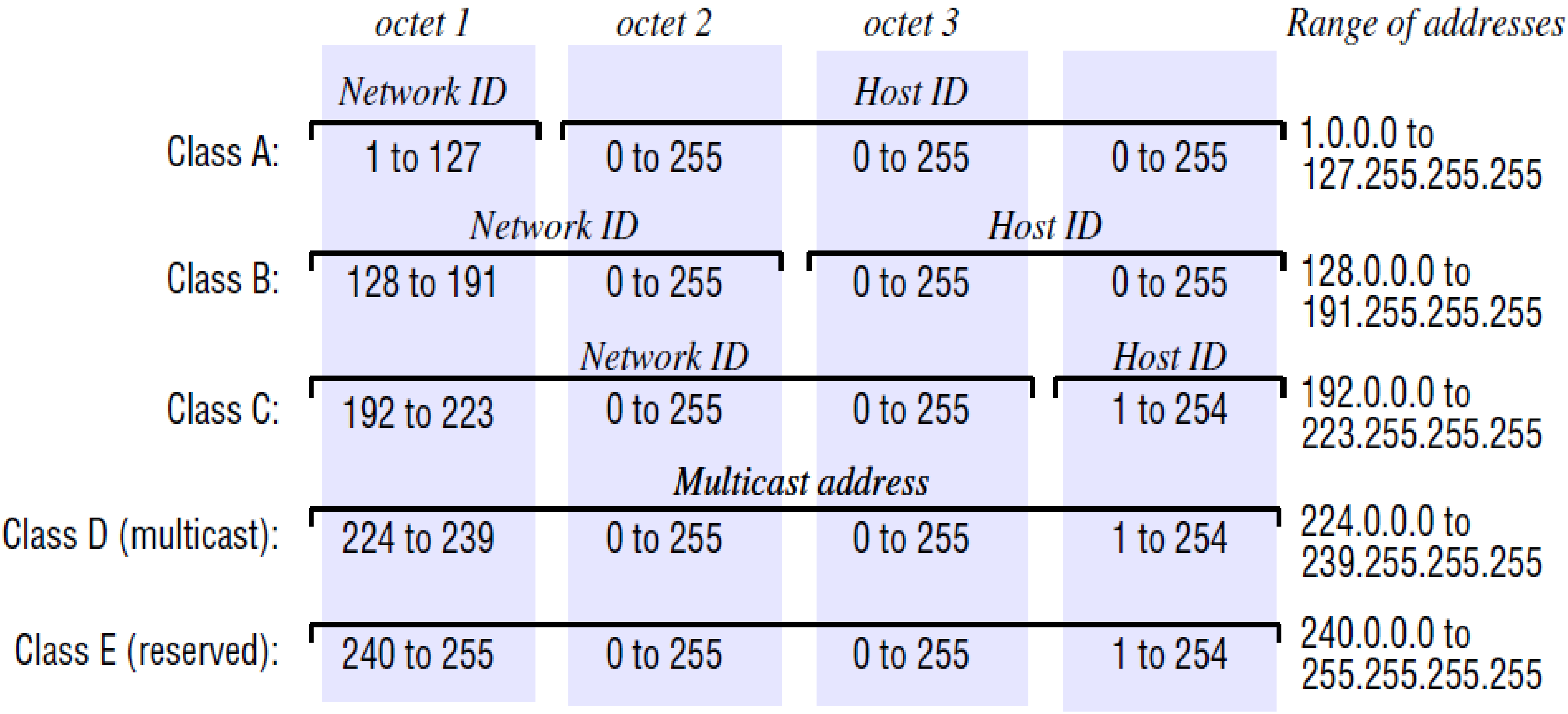

7
24
Class A: 0 Network ID Host ID
14
16
Class B: 1 0 Network ID Host ID
21
8
Class C: 1 1 0 Network ID Host ID
28
Class D (multicast): 1 1 1 0 Multicast address
28
Class E (reserved): 1 1 1 1 unused

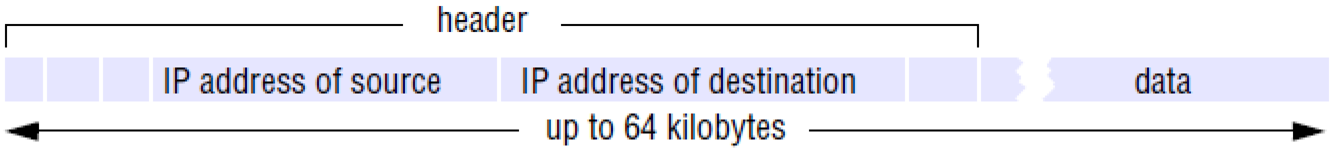
Figure 3.16 Decimal representation of Internet addresses



	octet 1	octet 2	octet 3		Range of addresses
	Network ID		Host ID		
Class A:	1 to 127	0 to 255	0 to 255	0 to 255	1.0.0.0 to 127.255.255.255
	Network ID		Host ID		
Class B:	128 to 191	0 to 255	0 to 255	0 to 255	128.0.0.0 to 191.255.255.255
		Network ID		Host ID	
Class C:	192 to 223	0 to 255	0 to 255	1 to 254	192.0.0.0 to 223.255.255.255
		Multicast address			
Class D (multicast):	224 to 239	0 to 255	0 to 255	1 to 254	224.0.0.0 to 239.255.255.255
Class E (reserved):	240 to 255	0 to 255	0 to 255	1 to 254	240.0.0.0 to 255.255.255.255

3.4.2 The IP protocol

Figure 3.17 IP packet layout



The IP protocol transmits datagrams from one host to another, if necessary via intermediate routers. Fig 3.17 shows its components

IP provides a delivery service that is described as offering *unreliable* or *best-effort* delivery semantics, because there is no guarantee of delivery. Packets can be lost, duplicated, delayed or delivered out of order, but these errors arise only when the underlying networks fail or buffers at the destination are full.

The only checksum in IP is a header checksum, which is inexpensive to calculate and ensures that any corruptions in the addressing and packet management data will be detected. There is no data checksum, which avoids overheads when crossing routers, leaving the higher-level protocols (TCP and UDP) to provide their own checksums – a practical instance of the end-to-end argument.

The IP layer must also insert a ‘physical’ network address of the message destination to the underlying network. It obtains this from the address resolution module in the Internet network interface layer. For example, if the underlying network is an Ethernet, the address resolution module converts 32-bit Internet addresses to 48-bit Ethernet addresses.

Address Resolution Protocol (ARP)

It uses dynamic enquiries in order to operate correctly when computers are added to a local network but exploits caching to minimize enquiry messages.

The ARP module on each host maintains a cache of (*IP address*, *Ethernet address*) pairs that it has previously obtained.

If the required IP address is in the cache, then the query is answered immediately. If not, then ARP transmits an Ethernet broadcast packet (an ARP request packet) on the local Ethernet containing the desired IP address.

Each of the computers on the local Ethernet receives the ARP request packet and checks the IP address in it to see whether it matches its own IP address. If it does, an ARP reply packet is sent to the originator of the ARP request containing the sender’s Ethernet address; otherwise the ARP request packet is ignored.

IP spoofing

We have seen that IP packets include a source address – the IP address of the sending computer. This, together with a port address encapsulated in the data field (for UDP and TCP packets), is often used by servers to generate a return address.

Unfortunately, it is not possible to guarantee that the source address given is in fact the address of the sender. A malicious sender can easily substitute an address that is different from its own. This loophole has been the source of several well-known attacks, including the distributed denial of service attacks .

The method used was to issue many *ping* service requests to a large number of computers at several sites (ping is a simple service designed to check the availability of a host). These malicious ping requests all contained the IP address of a target computer in their sender address field.

The ping responses were therefore all directed to the target, whose input buffers were overwhelmed, preventing any legitimate IP packets getting through.

3.4.3 IP routing

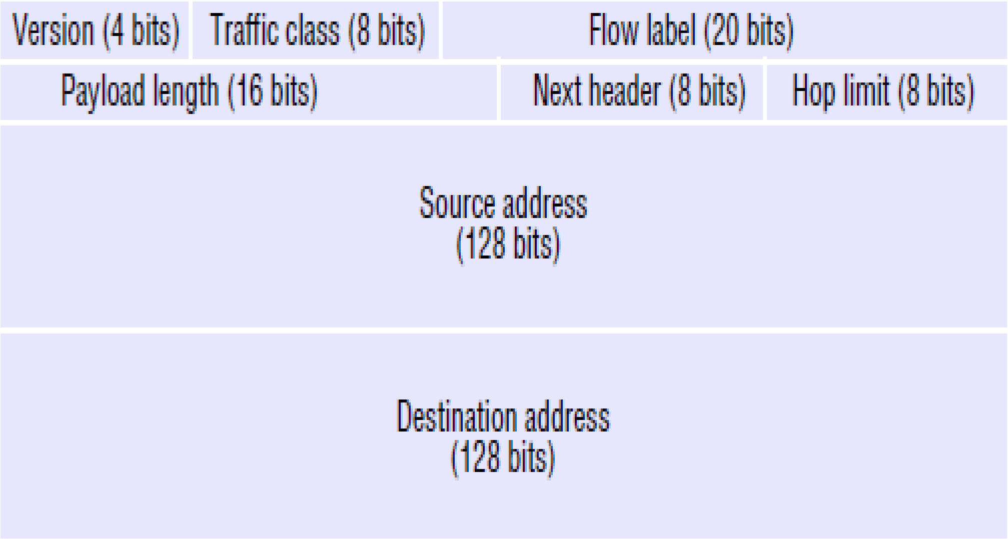
The IP layer routes packets from their source to their destination. Each router in the Internet implements IP-layer software to provide a routing algorithm.

Backbones • The topological map of the Internet is partitioned conceptually into *autonomous systems* (ASs), which are subdivided into *areas*. The intranets of most large organizations such as universities and large companies are regarded as ASs, and they will usually include several areas. Every AS in the topological map has a *backbone* area. The collection of routers that connect non-backbone areas to the backbone and the links that interconnect those routers are called the backbone of the network.

Routing protocols • RIP-1, the first routing algorithm used in the Internet, is a version of the distance-vector algorithm was developed from it to accommodate several additional requirements, including classless and erdomain routing, better multicast routing and the need for authentication of RIP packets to prevent attacks on the routers.

3.4.4 IP version 6 and 4

Figure 3.19 IPv6 header layout

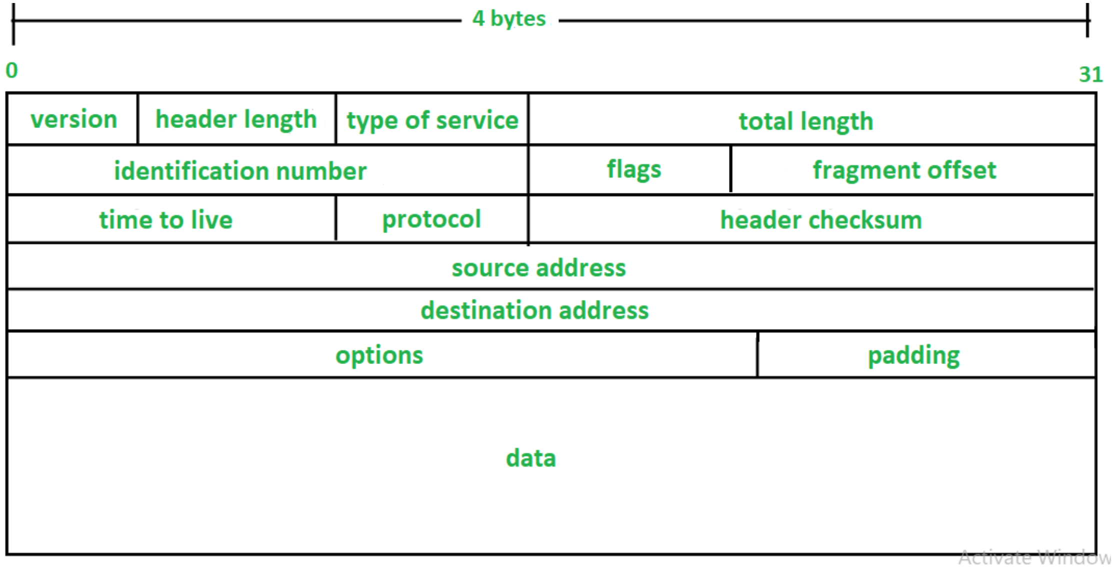


Version (4 bits) | Traffic class (8 bits) | Flow label (20 bits)

Payload length (16 bits) | Next header (8 bits) | Hop limit (8 bits)

Source address (128 bits)

Destination address (128 bits)



version | header length | type of service | total length

identification number | flags | fragment offset

time to live | protocol | header checksum

source address

destination address

options | padding

data

IPv6 addresses are 128 bits (16 bytes) long.

The first 6 bits of the *traffic class* field can be used with the *flow label* or independently to enable specific packets to be handled more rapidly or with higher reliability than others. Traffic class values 0 through 8 are for transmissions that can be slowed without disastrous effects on the application.

Other values are reserved for packets whose delivery is time-dependent. Such packets must either be delivered promptly or dropped – late delivery is of no value. Flow labels enable resources to be reserved in order to meet the timing requirements of specific real-time data streams, such as live audio and video transmissions.

3.4.5 MobileIP

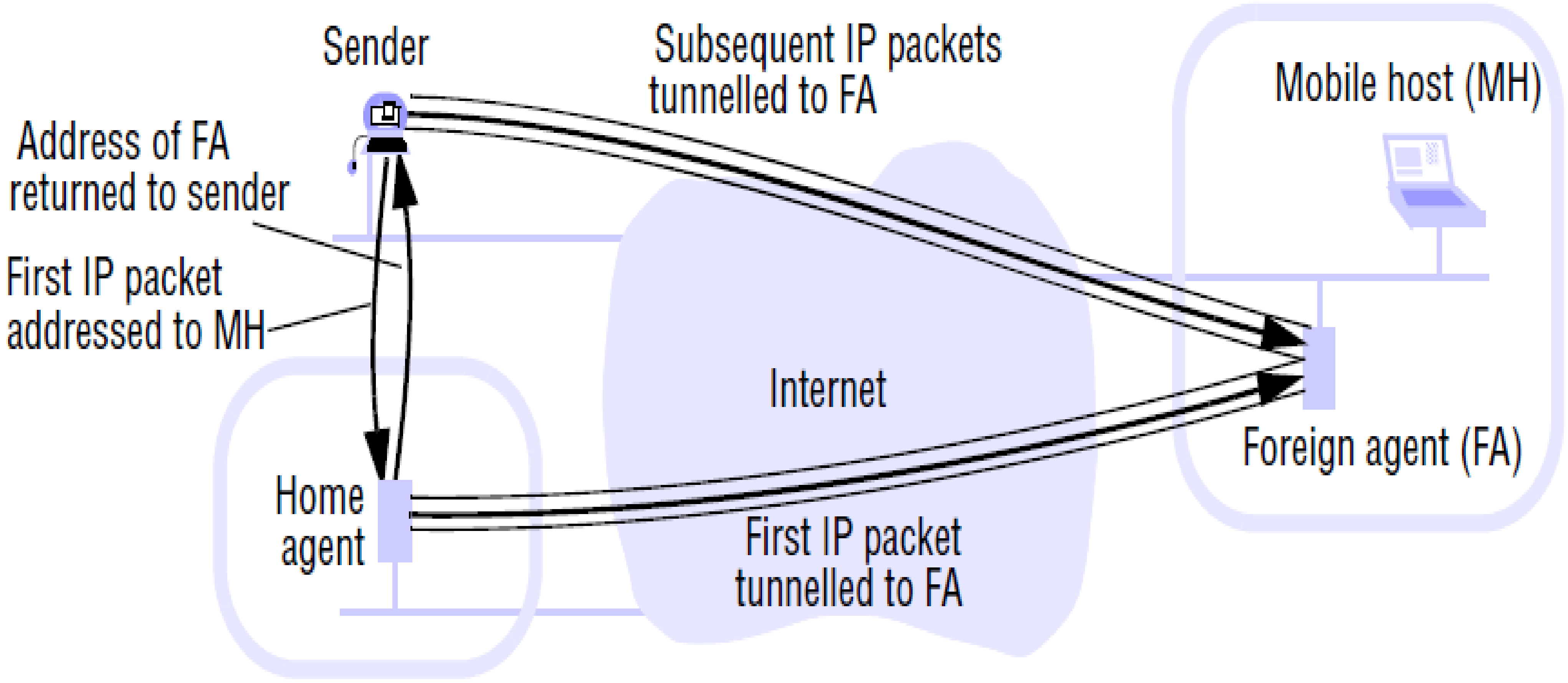
If a mobile computer is to remain accessible to clients and resource sharing applications when it moves between local networks and wireless networks, it must retain a single IP number, but IP routing is subnet-based. Subnets are at fixed locations, and the correct routing of packets to them depends upon their position on the network.

MobileIP is a solution for the latter problem. The solution is implemented transparently, so IP communication continues normally when a mobile host computer moves between subnets at different locations. It is based upon the permanent allocation of a normal IP address to each mobile host on a subnet in its ‘home’ domain.

When the mobile host is connected at its home base, packets are routed to it in the normal way. When it is connected to the Internet elsewhere, two agent processes take responsibility for rerouting. The agents are a *home agent* (HA) and a *foreign agent* (FA). These processes run on convenient fixed computers at the home site and at the current location of the mobile host. The HA is responsible for holding up-to-date knowledge of the mobile host’s current location.

When the mobile host arrives at a new site, it informs the FA at that site. The FA allocates a ‘care-of address’ to it – a new, temporary IP address on the local subnet. The FA then contacts the HA, giving it the mobile host’s home IP address and the care-of address that has been allocated to it.

Figure 3.20 The MobileIP routing mechanism


Sender
Subsequent IP packets tunnelled to FA
Mobile host (MH)
Address of FA returned to sender
First IP packet addressed to MH
Internet
Foreign agent (FA)
Home agent
First IP packet tunnelled to FA

3.4.6 TCP and UDP

The TCP layer includes additional mechanisms (implemented over IP) to meet the reliability guarantees. These are:

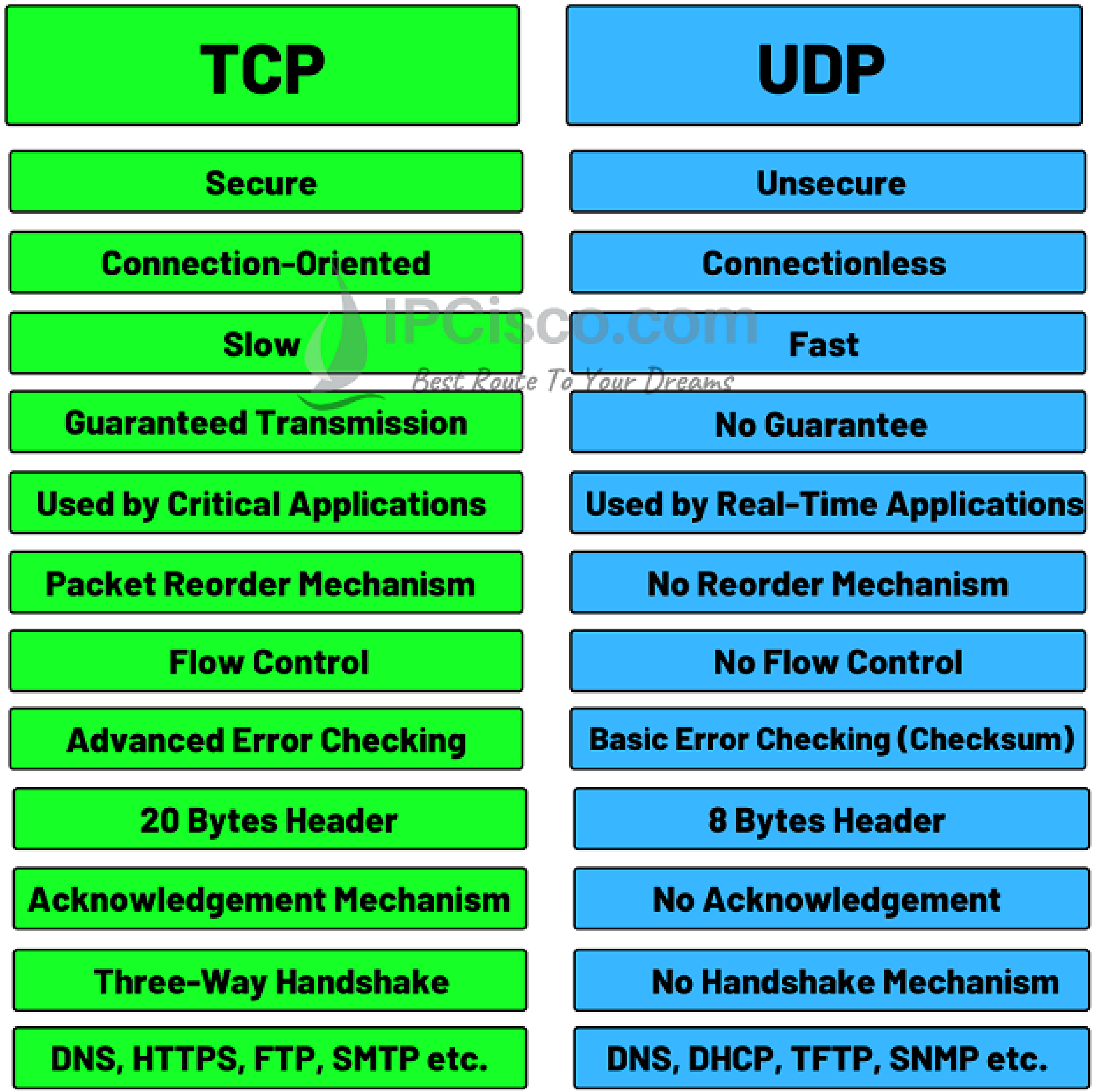
Sequencing:TCP sending process divides the stream into a sequence of data segments and transmits them as IP packets. A sequence number is attached to each TCP segment.

Flow control:The sender takes care not to overwhelm the receiver or the intervening nodes.

Retransmission: If any segment is not acknowledged within a specified timeout, the sender retransmits it.

Buffering:The incoming buffer at the receiver is used to balance the flow between the sender and the receiver.

Checksum: Each segment carries a checksum covering the header and the data in the segment. If a received segment does not match its checksum, the segment is dropped



TCP	UDP
Secure	Unsecure
Connection-Oriented	Connectionless
Slow	Fast
Guaranteed Transmission	No Guarantee
Used by Critical Applications	Used by Real-Time Applications
Packet Reorder Mechanism	No Reorder Mechanism
Flow Control	No Flow Control
Advanced Error Checking	Basic Error Checking (Checksum)
20 Bytes Header	8 Bytes Header
Acknowledgement Mechanism	No Acknowledgement
Three-Way Handshake	No Handshake Mechanism
DNS, HTTPS, FTP, SMTP etc.	DNS, DHCP, TFTP, SNMP etc.