~DRAFT~

CHAPTER 4: Network layer

Table of Contents

1. Introduction

The transport layer is able to provide process-to-process communication because it is supported by the network's host-to-host communication services, in turn supported by data-link layer services. The network layer must solve the problem of providing to the higher layers an abstraction of a channel that hides all the complexity of the underlying sub-

networks from the transport layer and ensures that information can be exchanged between hosts connected to different types of sub-networks or within sub-networks.

In this Chapter we show how this is done by discussing the functions of the network layer, see a generic view of the internal functioning of a router, discuss the two main algorithms for network routing (link state and distance vector routing algorithms), inter- and intra-Autonomous System (AS) routing and a few of the specific protocols, such as Border Gateway Protocol (BGP). We discuss the main two packet switching techniques of Datagram and Virtual Circuit packet switching. We also look at Internet Protocol (IP) versions 4 and 6 and the related IP protocols of Dynamic Host Configuration Protocol (DHCP), Network Address Translation (NAT) and Internet Control Message Protocol (ICMP).

2. Context

In the OSI model the network layer lies between the data-link layer below and the transport layer above. While we have seen that the transport layer provides an end-to-end service, with the protocols being implemented at the end-point hosts, the network layer involves the processes that occur within the network between the hosts to route and forward packets along the path from source to destination.

3. Objectives of Chapter

By the end of this Chapter you should be able to:

- Describe the functions of the networking layer
- Understand the difference between forwarding and routing
- Know the main components of a router and their functions
- Know the two main point-to-point routing algorithms

- Describe how link state algorithm works
- Describe how distance vector algorithm works
- Name some services offered by the network layer
- Describe the two connection service models in the network layer (Datagram and Virtual Circuit models), and compare and contrast their advantages and disadvantages
- Give some advantages of IPv6 over IPv4
- Know the different parts of an IPv4 address and what information can be gleaned from the address
- Describe the differences between inter-AS and intra-AS routing and name some protocols used for each
- Discuss in general the functions of DHCP, NAT and ICMP

4. Functions

The main purpose of the network layer is to move packets from a sender to the correct receiver in a multi-hop network by finding a route for the packets to follow and forwarding the packets. (In multi-hop networks, communication between two end hosts is carried out through a number of intermediate routers whose function is to relay information from one point to another.) The network layer takes s*egments* from the transport layer in a sending host, encapsulates each segment into a *datagram* (a network-layer packet), and sends the datagrams to a neighbouring router. The network layer in the receiving host receives the datagrams from the neighbouring router, extracts the transport-layer segments and delivers the segments up to the transport layer.

The main functions of the network layer router devices are:

- *Forwarding* packets arriving at a router's input link to the appropriate output link. Forwarding is a process that occurs within a single router, performed by the input ports, switching fabric and output ports as shown in Figure 1.
- *Routing*: determining the route or path packets will take over the network as they flow from a sender to a receiver. Routing is a network-wide process.

TABLE 1: EXAMPLE FORWARDING TABLE FOR DATAGRAM ROUTING

Prefix	Interface
A	0
C	1
B	2

Every router has a *forwarding table*, also known as the Forwarding Information Base (FIB) with information about destination addresses and the appropriate link interface on the router's output side. Usually only the prefix of the full destination address is stored in the FIB. A small example is given in Table 1 above. A router forwards a packet by examining the value of a field in the arriving packet's header, and then using this header value to index into the router's forwarding table. The value stored in the forwarding table entry for that header indicates the router's outgoing link interface to which that packet is to be forwarded. Depending on the network-layer protocol, the header value could be the destination address of the packet or an indication of the connection to which the packet belongs. A routing algorithm does the job of determining the values that are inserted into the routers' forwarding tables.

5. The architecture of a router

Figure 1 shows the main components of a generic router. Note that the block numbers in the input ports of Figure 1 correspond to the functions given below.

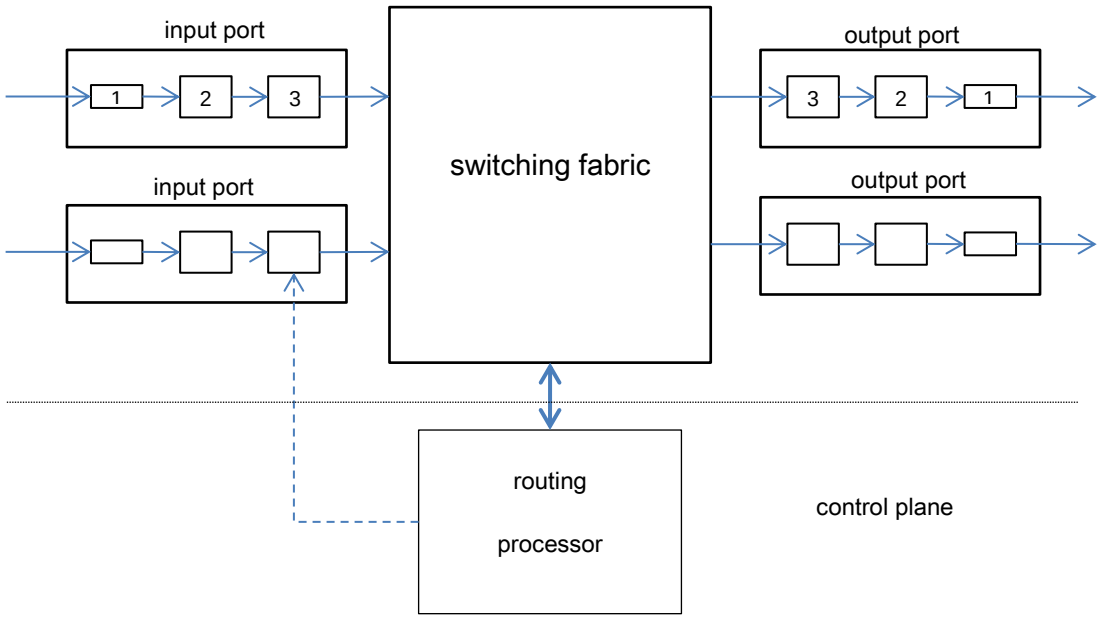

FIGURE 1: ARCHITECTURE OF A ROUTER

The following functions are performed at the *input ports* and in reverse order at the *output ports*, as indicated by the block numbers at the input and output ports in Figure 1:

1. Terminate an incoming physical link at the router (physical layer function)
2. Header processing (packet de-capsulation and encapsulation) and perform implemented data-link layer protocol functions (link layer function)
3. At the input port, perform the forwarding table lookup that determines to which output port an arriving packet will be forwarded via the switching fabric; and at both input and output ports queuing of datagrams and buffer management.

The *switching fabric* connects the router's input ports to its output ports.

The *output ports* store packets received from the switching fabric and transmit these packets on the outgoing links by performing the necessary link layer and physical layer functions.

The *routing processor* executes the routing algorithms, maintains routing tables and attached link state information, computes the forwarding table for the router and performs the network management functions. A shadow copy of the forwarding table computed by the routing processor is typically stored at each input port, enabling forwarding decisions to be

made locally at each input port, without invoking the centralised routing processor on a per-packet basis and preventing it from becoming a bottleneck.

Queuing of packets may occur at both the input and output ports. If the switching fabric has an operating rate slower than the arrival of packets on all lines, packets will be queued at the input port. In other words, if packets arrive faster than the switching fabric can process them, packets are queued. Once the buffer at the input port is full packets will be lost. If the rate at which the switching fabric can forward packets to the output ports is greater than the ports can transmit packets, queuing occurs at the output port. If this buffer is full packet loss of newly arriving packets will occur. The decision of which packet to forward next is performed by the packet scheduler. The decision may be made on a first in first out basis (FIFO), weighted fair queuing (WFQ) where the outgoing link is shared fairly among the different end-to-end connections, or any of an array of others.

6. Virtual Circuit vs Datagram networks

Virtual Circuit (VC) architecture is one that uses connections at the network layer to transmit data in such a way that it appears as though there is a dedicated physical layer link between the source and destination end systems. Being connection oriented means that signalling is required during a connection establishment phase and data is delivered in the correct order. In a VC network, a forwarding table in a router is modified whenever a new connection is set up through the router or whenever an existing connection through the router is torn down. This could easily happen at a microsecond timescale in a backbone, tier-1 router

A Virtual Circuit consists of:

- a path (that is, a series of links and routers) between the source and destination hosts
- VC numbers, one number for each link along the path
- entries in the forwarding table in each router along the path. A packet belonging to a virtual circuit will carry a VC number in its header. Because a virtual circuit may have a different VC number on each link, each intervening router must replace the VC

number of each traversing packet with a new VC number. The new VC number is obtained from the forwarding table.

In a VC network, the network's routers must maintain connection state information for the ongoing connections. Specifically, each time a new connection is established between hosts, a new connection entry must be added to the forwarding table of each router along the path and each time a connection is released, an entry must be removed from the table. Note that even if there is no VC-number translation, it is still necessary to maintain connection state information that associates VC numbers with output interface numbers. The issue of whether or not a router maintains connection state information for each ongoing connection is a crucial one.

The messages that the end systems send into the network to initiate or terminate a VC, and the messages passed between the routers to set up the VC (that is, to modify connection state in router tables) are known as signalling messages

Some advantages of VC packet switching

- Bandwidth reservation during the connection establishment phase is supported, making guaranteed Quality of Service (QoS) possible. For example a constant bit rate QoS class may be provided that emulates circuit switching.
- Less overhead is required. Only a small virtual channel identifier (VCI) is required in each packet. Routing information is only transferred to the network nodes during the connection establishment phase, packets are not routed individually and complete addressing information is not provided or required in the header of each packet.
- The network nodes are faster and have higher capacity in theory, since they only perform routing during the connection establishment phase, while connectionless network nodes perform routing for each packet individually.

In a *datagram network*, packets to be transmitted are stamped with the address of the destination end system and sent into the network. Each router through which the packet passes uses the packet's destination address to forward the packet to the next router. Each

router has a forwarding table that maps destination addresses to link interfaces; when a packet arrives at the router, the router uses the packet's destination address to look up the appropriate output link interface in the forwarding table. The router then intentionally forwards the packet to that output link interface. In a datagram network the forwarding tables are modified by the routing algorithms, which typically update a forwarding table with regularity on the order of a few minutes. Because forwarding tables in datagram networks can be modified at any time, a series of packets sent from one end system to another may follow different paths through the network and so may arrive out of order.

7. Routing algorithms

Routing algorithms may be executed in either a centralised or decentralised manner. Centralised routing means that the algorithm is run on a central site/router and routing information is then downloaded to each of the routers in the network. The disadvantage of this technique in the case that one of the network routers performs the algorithm, is that a single router in the network carries a great computational burden. Also, if the centralised router fails for some reason then the whole network essentially "crashes". Usually an algorithm executed in a centralised way will also be *global* – that is, executed with a complete global knowledge of the network.

In decentralised routing each router carries out a part of the routing algorithm. Routers may exchange control messages to interoperate. Routers may only have information about the routers it is directly connected to but not know about every router in the network or have complete information about the costs of all network links.

Routers are organised into sub-networks known as Autonomous Systems (ASs), each of which consists of a set of routers typically under the same administrative control of an organisation or company, which share routing information with each other and use a common routing protocol. This routing protocol uses an *intra-AS* routing algorithm since it is only used within the AS. Between different ASs, routers run *inter-AS* routing protocols that require one or more router in each AS to act as the gateway to other ASs.

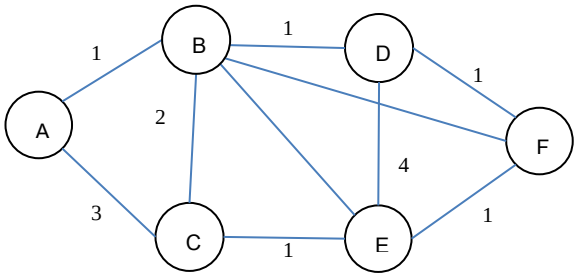

FIGURE 2: GRAPH ABSTRACTION SHOWING ROUTER NODES A TO F AS GRAPH VERTICES AND EDGES AS POSSIBLE PATHS

Figure 2 shows a graph representation of a network with router nodes A-F, and link costs indicated on the edges of the graph. A graph model denoted $G = (V, E)$ is usually used in formulating routing algorithms. The nodes/vertices V in the graph G represent routers and the edges E connecting these nodes represent the physical links between these routers. Each link can be assigned a cost to transmit a packet over that link as indicated on the edges of the graph in Figure 2, so the problem of routing a packet from source (say router A) to destination (say router B) is a matter of solving the computing problem of finding the (or a) least costly path. A *path* is a sequence of nodes through which a packet passes on its way from source to destination.

The main two routing algorithms for unicast applications in a Datagram network are Link-State and Distance Vector. Unicast means there is only one sender and one receiver per packet.

a. Link State Routing

In link-state routing, a network is modelled as a *directed weighted graph*. The algorithm has upfront knowledge of, and uses as input, the network topology and all link costs. Because the algorithm has complete information about the network it is known as a global routing algorithm. A positive weight is associated to each directed edge from the information. In practice this is accomplished by having each node broadcast link-state packets to all other nodes in the network, with each link-state packet containing the identities and costs of its attached links, so that all nodes have an identical and complete view of the network and can compute the same set of least-cost paths. Usually, the same weight is associated to the two

directed edges that correspond to a physical link (i.e. $R1 \rightarrow R2$ and $R2 \rightarrow R1$). However, the link state protocol does not require this. The actual shortest path computation may then be performed by an algorithm such as Dijkstra's algorithm or other.

Open Shortest Path First (OSPF) is a well-known link state protocol for routing within an Autonomous System. OSPF floods link-state information to all other routers in the AS, as opposed to just neighbours. Each router constructs a complete graph of the AS with the link cost information configured by the network administrator such that the protocol can achieve certain traffic engineering goals (example minimum hop routing, encourage using higher bandwidth routes).

b. Distance Vector routing

The name Distance-Vector (DV) algorithm is named for the fact that routes are advertised as vectors of (distance, direction), where distance is defined in terms of a link metric and direction is defined in terms of the next-hop router. The distance-vector routing algorithm is iterative, asynchronous and distributed and makes routing decisions based on the number of hops between source and destination. It is distributed in that each node receives some information from one or more of its directly attached neighbours, performs a calculation independently, and then distributes the results of its calculation back to its neighbours (not the whole network). Since all nodes perform part of the algorithm and none have a complete view of the network DV algorithm is decentralised. At the start it is assumed that each node knows the cost of the link to each of its directly connected neighbours. Each router creates a distance-vector table with the known costs to next hop routers. A link that is down is assigned an infinite cost. Each node maintains a table such as the example in Table 2.

TABLE 2: EXAMPLE DISTANCE VECTOR TABLE MAINTAINED AT EACH ROUTER

Destination	*Cost*	*Next hop*
A	1	A
C	1	C
D	2	C

E	2	A
F	2	A
G	3	A

The algorithm is iterative since it continues this process of receiving information updates from neighbours, calculating new distance vectors from the information, updating its table and sharing new distance vector information until no more new information can be exchanged. The only information a node has is the costs of the links to its directly attached neighbours and information it receives from these neighbours. As long as all the nodes continue to exchange their distance vectors in an asynchronous fashion, each individual cost estimate converges to the actual cost of the least-cost path between the same two points.

The Routing Information Protocol (RIP) is one of the earliest Distance-Vector protocols for routing within a network, which is rarely used today. Other examples of DV protocols are Xerox Networking System's XNS RIP, the Border Gateway Protocol (BGP), Cisco's Internet Gateway Routing Protocol (IGRP) and AppleTalk's Routing Table Maintenance Protocol (RTMP).

Table 3 compares the two kinds of routing algorithms.

TABLE 3: COMPARISON OF LS AND DV ROUTING ALGORITHMS

Attribute	*LS*	*DV*
Message complexity	LS requires each node to know the cost of each link in the network. The new link cost must be sent to all nodes on every change	DV requires message exchanges between directly connected neighbours at each iteration. When link costs change, results of the changed link cost will only be propagated if the new link cost results in a changed least-cost path for one of the nodes attached to that link
Robustness	Should a link fail, a router could broadcast an incorrect cost for one of	A node can advertise incorrect least-cost paths to any or all destinations

	its attached links (but no others). A node could also corrupt or drop any packets it received as part of an LS broadcast. But an LS node is computing only its own forwarding tables; other nodes are performing similar calculations for themselves. This means route calculations are somewhat separated under LS, providing a degree of robustness.	so an incorrect node calculation can be diffused through the entire network under DV, which may cause disconnections for extended periods.

In the DV algorithm, each node talks to only its directly connected neighbours, but it provides its neighbours with least-cost estimates from itself to all the nodes (that it knows about) in the network. In the LS algorithm, each node talks with all other nodes (via broadcast), but it tells them only the costs of its directly connected links. It is up to the network administrator to select whether a DV or LS routing protocol is used and which one based on goals that are relevant to the organisation. Often the brand of router purchased will require the use of a specific protocol but not always. If compatibility with legacy equipment is a decision factor, this may limit the options. The selection should consider the anticipated network traffic characteristics of the organisation using the network; efficiency in bandwidth, memory and CPU usage of the protocol; the speed and ability to adapt to changes and possibly security and authentication capabilities. Ease of configuration and management and cost are other important factors to consider. The metrics supported by the specific algorithm must be matched to the goals identified by the administrator, for example bandwidth, delay, network traffic, reliability and hop count may have varying levels of importance in different situations.

So far we have focused on the situation where there is only one source and one destination (unicast). Multicast, broadcast and even anycast routing algorithms also exist. Broadcast packets are usually not forwarded and routed because as routers create broadcast domains. To route a broadcast packet, one way is for the router to send it to each host one by one. In this case, the router creates multiple copies of the single data packet with different

destination addresses. All packets are sent as unicast but because they are sent to all, it simulates broadcasting. Disadvantages of this broadcast routing method are that it consumes a lot of bandwidth and the router must know the destination address of each node. Another way for a router to broadcast packets is to flood those packets out of all interfaces, which may result in duplicated packets. In anycast packet forwarding multiple hosts can have same logical address. When a packet destined to this logical address is received, it is sent to the host which is nearest in routing topology. Anycast routing is done with help of Domain Name System (DNS) server.

In very large networks both LS and DV would experience problems. For example, the public internet consists of over a billion hosts[1]. Storing routing information at each of these hosts would require enormous amounts of memory. A distance-vector algorithm that iterated among such a large number of routers would also never converge, certainly not in an acceptable time period. The overhead required to broadcast LS updates among all of the routers in the public Internet would leave no bandwidth left for sending data packets!

For this reason, as well as the practical consideration of organisations requiring a level of autonomy and opacity over the routing within its own network, routers are grouped into Autonomous Systems. In each such Autonomous System the routers are controlled by one organisation or administrator, share the same routing algorithm which has information only about the routers within that network domain. To interconnect Autonomous Systems, one or more routers in a system will act as the *gateway* to forward packets to and from other autonomous systems.

If there is only one gateway router with one outward link to other Autonomous Systems routing between ASs is simple. For the case with more than one exit point from one AS to the next, an inter-AS routing protocol is used. Each AS learns which destinations are reachable via each egress point and propagates this reachability information to all the routers within the AS, so that each router can configure its forwarding table to handle external-AS destinations. If a destination can be reached through more than one gateway

[1] As of 2016 according to https://www.statista.com/statistics/264473/number-of-internet-hosts-in-the-domain-name-system/

and neighbouring AS, the sending router may then send the packet to the gateway router that has the smallest router-to-gateway cost. This is known as *hot-potato routing*.

c. Border Gateway Protocol

The Border Gateway Protocol (BGP) has been called the *de facto* standard for inter-AS routing in today's internet. BGP allows an AS to obtain subnet reachability information from neighbouring ASs, propagate the reachability information to routers within the AS, use the information to determine good routes to subnets and advertise its existence to the rest of the internet.

8. Network layer services

A third possible (but not necessary) function in some network-layer architectures (for example ATM, frame relay, and MPLS) is *connection setup*. These architectures require the routers along the chosen path from source to destination to handshake with each other in order to set up state before network-layer data packets within a given source- destination connection can begin to flow. In the network layer, this process is referred to as *connection setup*.

The *network service model* defines the characteristics of end-to-end transport of packets between sending and receiving end systems. In the sending host, *when the transport layer passes a packet to the network layer*, specific services that could be provided by the network layer include guaranteed delivery (with no time guarantee) or guaranteed delivery with bounded delay, which not only guarantees delivery of the packet, but delivery within a specified host-to-host time delay bound (for example, within 100 ms)

Examples of services that could be provided to a *flow of packets* between a given source and destination include:

- Guaranteed in-order packet delivery at the destination.

- Guaranteed minimal bandwidth. As long as the sending host transmits packets at a rate below the specified bit rate, then no packet is lost and each packet arrives within a pre-specified host-to-host delay (for example, within 40 ms).
- Guaranteed maximum jitter, which guarantees that the amount of time between the transmission of two successive packets at the sender is equal to the amount of time between their receipt at the destination (or that this spacing changes by no more than some specified value).
- Security services. Using a secret session key known only by a source and destination host, the network layer in the source host could encrypt the payloads of all datagrams being sent to the destination host. The network layer in the destination host would then be responsible for decrypting the payloads. With such a service, confidentiality would be provided to all transport-layer segments (TCP and UDP) between the source and destination hosts. In addition to confidentiality, the network layer could provide data integrity and source authentication services.

9. The Internet Protocol

a. The IP Datagram

The IP datagram includes some fields that are key to its various functions. These include the version number (indicating IPv4 or IPv6), header length, type of service (for example real-time or non-real-time), datagram length, time to live, header checksum to help detect bit errors source and destination IP addresses, amongst other fields in addition to the data payload.

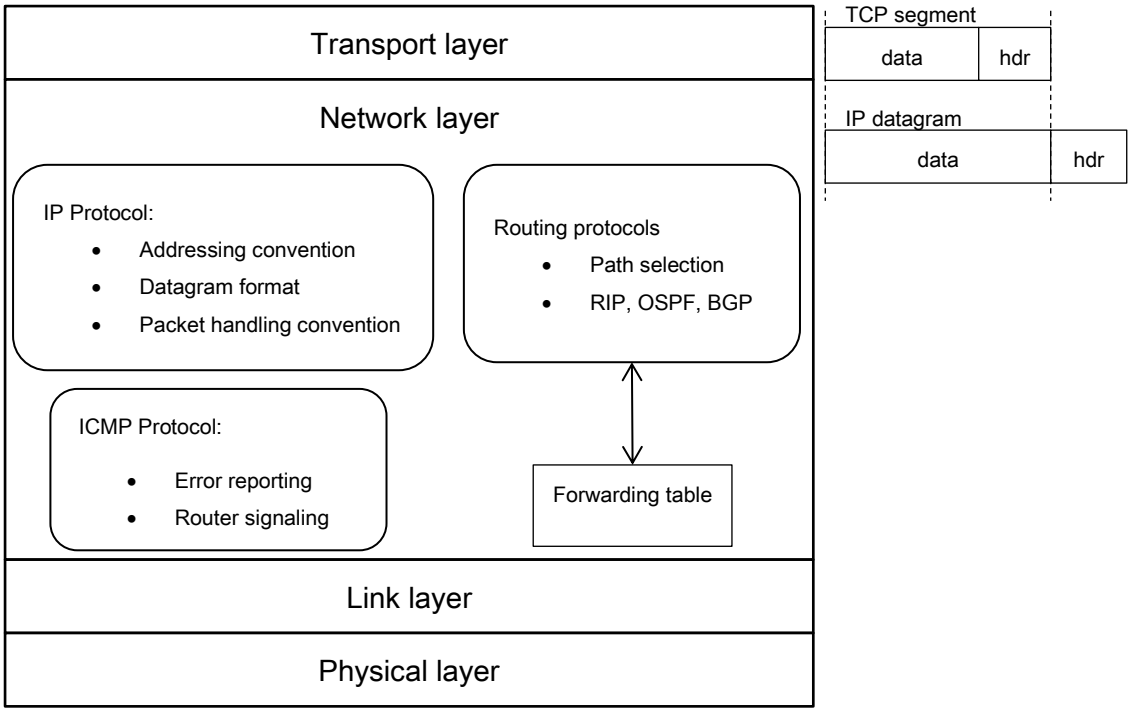

FIGURE 3: THE IP NETWORK LAYER SHOWING A DATAGRAM RELATED TO A TCP SEGMENT

Link layer protocols are each able to carry different sized network layer datagrams. For example, Ethernet frames can carry up to 1500 bytes of data but frames for wide area links will only be able to carry up to 576 bytes of data. The value that indicates the maximum amount of data that can be carried by a link layer frame is the Maximum Transmission Unit (MTU). This value limits the length of an IP datagram. If an IP datagram is larger than the MTU of the outgoing link, the router fragments the datagram and encapsulates each fragment in a link layer frame to be transmitted.

Between a sender and destination numerous different link layer protocols may be in use along the way. The sending host stamps each outgoing IP datagram fragment with the identification number of the original datagram (before fragmentation) and source and destination addresses, incrementing the identification number of each datagram sent, a flag, and fragmentation offset. The destination can use these values to determine which of the fragments it receives belong to the same original datagram. The offset field is used to specify where the fragment fits within the original IP datagram. The last fragment in a series

of fragments of the same datagram has the flag bit set to 0 while all the others have the flag set to 1. This enables the destination to know when it has received the final fragment. The IP network layer at the destination reconstructs the datagram before sending it up to the transport layer. The process of fragmentation introduces a vulnerability to denial of service attacks.

b. IPv4

An IP address is a number that uniquely identifies a host on a TCP/IP network. In IPv4 addressing each IP address has 4 bytes (32 bits). The IP address is usually written as 4 decimal values (each representing a byte or octet) separated by a period, such as 192.168.3.1, called dotted-decimal notation. To obtain the byte representation each decimal is converted to its binary equivalent and written with spaces in between the octets so the IP address above (192.168.3.1) in byte representation is

11000000 10101000 00000011 00000001

In order for a router to be able to route a packet from a host in an AS to another host in a separate AS, the first part of the destination address is reserved to represent the network to which the packet must be routed. The second part is the host's unique address. Generally the first three octets (24 most significant bits) gives the *network prefix* representing the subnetwork (subnet) to which the host belongs and the last octet is the unique address of the host connection on its subnet (also called the *rest field*). In our example 192.168.3.0 is the network address and 0.0.0.1 is the host address. All hosts belonging to the same network share a common network prefix. However, the parts of the IP address that are used as the network and host addresses are not fixed across all networks. The *subnet mask* is another IP address that provides the information the routers need to identify and separate out the network and host addresses. The number of least significant bits that are 0 in the subnet mask gives the number of bits occupied by the host address portion of the destination address, and those that are 1 are the network prefix portion. In binary, for the example above, the subnet mask would thus be

11111111 11111111 11111111 0000000

or 255.255.255.0 in dotted-decimal notation, which shows that the first 24 bits are the network prefix and the last 8 bits are the host suffix. The network address can be separated out by taking the bitwise logical AND of the binary representations of the subnet mask and the destination IP address. Internet RFC 1878 describes the valid subnets and subnet masks that can be used on TCP/IP networks. The address 255.255.255.255 is the IP broadcast address, which means a datagram with that destination address will be sent to all hosts on the subnet. This address cannot be assigned for other purposes.

Subnetting is the process of dividing a network into smaller subnets. A TCP/IP network can be subnetted, by a system administrator to streamline the administration in a way that matches divisions within the organisation, for example a company with departments in different cities may subdivide addresses into separate subnets for departments. The administrator could divide the network into subnets by using a subnet mask that makes the network address larger and the possible range of host addresses smaller. For example, a subnet such as 255.255.255.192 could be used, which gives four networks of 64 hosts each. Using this subnet mask the 192.168.3.0 network then becomes the four networks given in Table 4 below. Binary host addresses with all ones or all zeros are invalid. NOTE: a subnet mask does not work like an IP address, nor does it exist independently from the IP addresses.

TABLE 4: EXAMPLE SUBNET NETWORK ADDRESSES AND VALID HOST ADDRESSES PER SUBNET

Network address	Valid host addresses
192.168. 3.0	192.168.3.1-62
192.168. 3.64	192.168. 3.65-126
192.168. 3.128	192.168.123.129-190
192.168. 3.192	192.168.123.193-254

The Classless Inter-domain Routing (CIDR) notation is often used to represent an IP address and the associated routing prefix. The CIDR has the form 192.168.1.0/24. The number after the slash (/) character indicates that 24 bits are allocated for the network prefix, and the remaining 8 bits are reserved for host addressing. The unique host address may be

assigned automatically with the Dynamic Host Configuration Protocol (DHCP) by a network server or manually by an administrator. The (much) older alternative to CIDR is *classful* addressing. Networks were divided into classes A, B and C: A having 8-bit subnet addresses, B 16-bit and C 24-bit subnet addresses. This had the requirement that the subnet part of an IP consist of exactly 1, 2 or 3 bytes according to the class. It was wasteful causing a shortage of available addresses and was done away with after the introduction of CIDR in 1993.

c. IPv6

IP version 6 was developed by the Internet Engineering Task Force, and presented in Internet standard document RFC 2460, primarily to combat the rapidly worsening problem of running out of IPv4 addresses. IPv6 uses 128 bit addresses, allowing many orders of magnitude more devices to be connected to the internet than the 32 bit addresses of IPv4. IPv6 addresses are represented as eight groups of four hexadecimal digits with the groups being separated by colons, for example 2001:0db8:0000:0042:0000:8a2e:0370:7334. The last unassigned top-level address blocks of 16 million IPv4 addresses were allocated in February 2011 by the Internet Assigned Numbers Authority (IANA) to the five Regional Internet Registries (RIRs). The African Network Information Centre (AFRINIC) is the sole RIR that is still using the normal protocol for distributing IPv4 addresses. The rest of the world's RIRs have reached a stage of address depletion where all the blocks they have not reserved for IPv6 transition have already been allocated.

The IPv6 protocol was not designed to be backwards compatible; consequently IPv4 and v6 are not compatible or interoperable and IPv4 and v6 networks essentially exist in parallel unless another transition mechanism is employed that allows interoperability. Network Address Translation (NAT)-Protocol Translation (RFC-2766) or NAT64 are two ways. To use NAT-PT dedicated hardware devices must be placed at the edge of the IPv6 network. Another transition mechanism is using a tunnelling protocol such as 6to4, 6in4, or Teredo where IPv6 packets are encapsulated into IPv4 packets. The 6to4 protocol does not require manual tunnel set-up but does use dedicated relay routers for forwarding. RFC 4213 Basic

Transition Mechanisms for IPv6 Hosts and Routers gives the technical basics for this. The 6over4 protocol is an elegant solution for interconnecting isolated IPv6 hosts in an IPv4 site where IPv6 multicast is implemented over IPv4 multicast and the IPv6 Neighbour Discovery mechanism is used by IPv6 nodes to configure themselves. Unfortunately, IPv4 multicast is not generally available on all networks, and there are scalability issues with this approach. It is ideal for small self-contained networks where multicasting is available. Another method is dual stack where the network hardware runs IPv4 and IPv6 simultaneously in parallel. This may require the network administrator to deploy dual-stack capable switches. By using Dual Stack Application Level Gateway, dual-stack servers are used as proxies to perform protocol translation with one proxy server per application (e.g. http, ftp, smtp). This has the advantage that very few IPv4 addresses are required since they are only needed for the proxies and not for all routers, and the protocol translation step may not be such a large price to pay in situations where firewalls and proxy servers already exist, which is the case in many LANs.

Features of IPv6:

- Larger address space
- Multicasting as part of the base specification (this was an optional feature in IPv4) and extended support for different implementations of multicasting. Multicast can also be used to broadcast so there is no reserved broadcast address in IPv6.
- Stateless address auto-configuration. Using the Neighbour Discovery Protocol via Internet Control Message Protocol version 6 (ICMPv6) router discovery messages, hosts can configure themselves automatically when connected to an IPv6 network. Stateful configuration with DHCP (discussed in Section d) version 6 is still supported as an alternative, or hosts may be configured manually using static methods.
- Network-layer security through IP Security (IPsec).
- Simplified 40-byte packet header. Rarely used fields in IPv4 have been moved to optional header extensions in IPv6.
- Simplified packet forwarding process. Less processing is done in the routers as some functionality has been moved to the edge of the network. The routers in IPv6 do not perform IP fragmentation. Instead the hosts are required to either perform path MTU

discovery, end-to-end fragmentation, or to send packets no larger than the default MTU, which is 1280 octets.

- Jumbograms, which are optional datagrams that exceed the payload size limit of 65 535 octets.

d. DHCP

The DHCP client-server protocol is used for the management and automation of network configuration of devices connected to an IP network. Once an organisation has obtained a block of addresses the network administrator must ensure individual IP addresses are assigned to all host and router interfaces. Usually this is done using the DHCP protocol. DHCP is an extension of an earlier network IP management protocol called Bootstrap Protocol (BOOTP) but DHCP is more advanced, and DHCP servers can handle BOOTP client requests if any BOOTP clients remain on a network segment. All common network devices including phones, PCs and other consumer gadgets support DHCP, which is built into the software of all common network operating systems.

DHCP allows a host to obtain an IP address automatically. It is possible for the network administrator to configure DHCP in such a way that a certain host receives the same IP address every time it connects to the network but if not, the host would be allocated a temporary address that will be different every time. DHCP is designed for use within a limited specific local area network (LAN). If network administrators want a given DHCP server to provide addressing to multiple subnets on a given network, they must configure DHCP relay services on the routers through which requests have to pass. Most simply each subnet will usually have a DHCP server but, if not, a DHCP *relay agent* that knows the address of a DHCP server for that network is required.

DHCP environments require a DHCP server to be set up with the appropriate configuration parameters for the given network. In small businesses a router may act as the DHCP server while in larger networks a single computer might act as the DHCP server. DHCP servers assign, then release and renew these addresses as devices leave and re-join the network.

DHCP parameters include the range of available IP addresses, the correct subnet masks, network gateway address, and domain name and domain name server address. The first step in configuring a DHCP server is to create the configuration file that stores the network information for the clients. This file is used to declare options and global options for client systems. Once the DHCP server is configured and started, DHCP client devices can automatically retrieve these settings from the DHCP servers as needed, provided they are configured to do so.

When a client is initialised for the first time after it is configured to receive DHCP information or when a new host enters a network it must find a DHCP server and initiate communications. It does this by broadcasting a DHCPDISCOVER message within a UDP packet to port 67, encapsulated into an IP datagram, to the broadcast destination IP address 255.255.255.255, with a source address of 0.0.0.0. Thereafter the DHCP server responds by broadcasting a DHCPOFFER message to all nodes on the subnet using the same broadcast address of 255.255.255.255. The offer message contains the transaction ID of the received discover message, the proposed IP address for the client, the network mask, and an IP address lease time — the amount of time for which the IP address will be valid – which may be of the order of hours or days. The client can choose between offers if it receives multiple DHCP server offers and respond to the chosen server with a DHCPREQUEST message. The source address in this message is still 0.0.0.0 since the client has not yet received verification from the server to start using the new address offered, and the destination for the packet is still 255.255.255.255 to indicate to the other DHCP servers that they can release their offered addresses and return them to their available pools (since another offer was accepted). The DHCP server responds to the DHCPREQUEST with a DHCPACK that confirm the requested parameters, thus completing the initialisation cycle. If the client has previously had a DHCP assigned IP address and it is restarted, the client will specifically request the previously leased IP address in a special DHCPREQUEST packet.

e. Network Address Translation

In general Network Address Translation (NAT) allows remapping of one IP address space into another by modifying network address information in IP datagram packet headers while they are in transit across a traffic routing device. NAT allows a single device, such as a router, to act as an agent between the Internet (or "public network") and a local (or "private") network. The router assigns a public address to hosts inside a private network or Local Area Network (LAN). This means that only a single, unique IP address is required to represent an entire group of computers. The main benefit of NAT is to limit the number of public IP addresses an organisation must use, for both economy and security purposes.

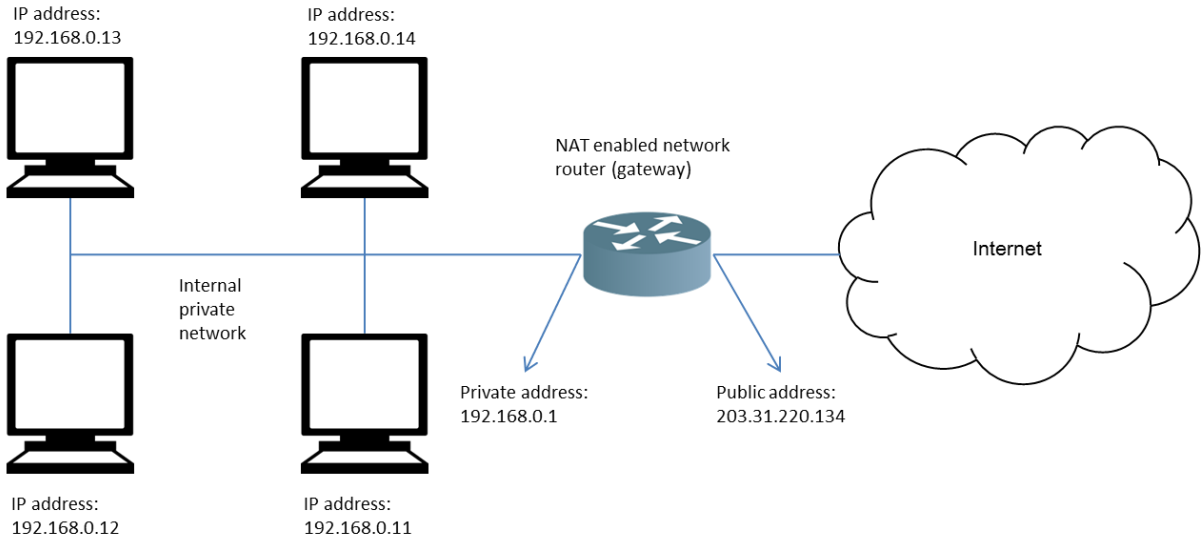

FIGURE 4: ILLUSTRATION OF THE NAT CONCEPT. THE NAT-ENABLED ROUTER MAPS ALL THE INTERNAL PRIVATE IP ADDRESSES TO A SINGLE OUTWARD-FACING PUBLIC ADDRESS

NAT is often used in private networks that use addresses in a private range. Private addressing is applicable for computers that only have to access resources inside the network, like workstations needing access to file servers and printers. However, to access resources outside the network, like the Internet, these computers have to have a public address in order for responses to their requests to return to them. The concept is illustrated in Figure 4. In the illustration there are four hosts (computers) in the internal network and one router that connects this network to the Internet. The Internet sees this internal network as only one device with the router's public IP address. NAT can also be used to limit access to the outside of the network. Computers requiring special access outside the network can be assigned specific external IP addresses using NAT, allowing them to communicate with

computers and applications that require a unique public IP address. In this case a firewall acts as the intermediary, and can control the session in both directions, restricting port access.

Essential to the NAT protocol, is the construction of a NAT translation table in the router on which NAT is run. The table records the port numbers and IP addresses of the private and applicable external interfaces. The port numbers are used to address the hosts in the internal network since the response from the external network will only have the NAT router's external IP address. Static NAT allows one-to-one mapping between local and global addresses. Dynamic NAT maps unregistered IP addresses to registered IP addresses from a pool of registered IP addresses. NAT with Port Address Translation (NATP), also known as NAT Overload and IP masquerading, takes a Static or Dynamic IP Address that is bound to the public interface of the gateway (PC, router or firewall appliance) and allows all computers within the private network to access the Internet. NATP uses a single public IP Address for the routing process and changes, in most cases, the Source or Destination port depending on whether it's an incoming or outgoing packet.

While end-to-end connectivity is generally regarded as a core principle of the Internet, NAT violates this concept. Hosts behind NAT-enabled routers do not have end-to-end connectivity and cannot participate in some Internet protocols. Services that require the initiation of TCP connections from the outside network, or some protocols using UDP, can be disrupted by the implementation of NAT. Unless the NAT employs specific extra techniques to support such protocols, incoming packets cannot reach their destination. Use of NAT complicates tunnelling protocols such as IP Security (IPsec). IP packets have a checksum in each packet header, which provides error detection only for the header. IP datagrams may become fragmented and it is necessary for NAT to reassemble these fragments to allow correct recalculation of the checksums and correct tracking of which packets belong to which connection. NAT modifies values in the headers of IP datagrams, which interferes with the integrity checks done by IPsec and other tunnelling protocols.

Many application protocols carry IP addresses in an application-level protocol. In such cases, an Application-Level Gateway (ALG) is needed to complete the translation if NAT is being used. Examples of such protocols are

- Many Internet Control Message Protocol (ICMP) packets such as "Destination Unreachable" carry embedded IP packets in ICMP payload. These require both address translation and checksum regeneration. (ICMP is discussed further in Section 9.f
- A File Transfer Protocol (FTP) ALG is needed to rewrite IP addresses carried by some FTP control commands. If the host making the request lies behind a simple NAT firewall, the translation of the IP address and/or TCP port number makes the information received by the server invalid
- Session Initiation Protocol (SIP), which controls many Voice over IP (VoIP) calls also relies on the initial IP address and TCP port number.
- Simple Network Management Protocol (SNMP) packets carry IP addresses that identify trap source and object instance. Dynamic NAT makes it impossible to uniquely identify hosts by IP address since public addresses are transient and shared. Remote management of private hosts can thus be impeded by NAT.
- DNS, responsible for domain name/IP address mapping, is impacted by NAT and requires an ALG.

f. Internet Control Message Protocol

For the Internet as we know it to work requires three main components in the network layer: the IP Protocol, the routing protocols, and finally the Internet Control Message Protocol (ICMP). ICMP enables hosts and routers to communicate network layer control information to each other, for example diagnostics, error reporting and simple queries. ICMP errors are directed to the source IP address of the originating packet.

ICMP messages are carried as IP payload inside IP datagrams so architecturally it lies above the IP layer and is not strictly part of it. The ICMP header contains type and code fields which together identify specific messages, and a checksum field for detection of errors

introduced during transmission. Some common type and code combination are listed in Table 5. For type 3 messages codes 0, 1, 4, and 5 may be received from a gateway and codes 2 and 3 may be received from a host.

TABLE 5: COMMON ICMP TYPE AND CODE DESCRIPTIONS

Type	Name	Code	Description
0	Echo Reply		(used to *ping*)
1	Unassigned		
2	Unassigned		
3	Destination Unreachable	0	Net Unreachable
3	Destination Unreachable	1	Host Unreachable
3	Destination Unreachable	2	Protocol Unreachable
3	Destination Unreachable	3	Port Unreachable
3	Destination Unreachable	4	Fragmentation Needed and Don't fragment was Set
3	Destination Unreachable	5	Source Route Failed
3	Destination Unreachable	6	Destination Network Unknown
3	Destination Unreachable	7	Destination Host Unknown
3	Destination Unreachable	8	Source Host Isolated
3	Destination Unreachable	9	Communication with Destination Network is Administratively Prohibited
3	Destination Unreachable	10	Communication with Destination Host is Administratively Prohibited
3	Destination Unreachable	11	Destination Network Unreachable for Type of Service
3	Destination Unreachable	12	Destination Host Unreachable for Type of Service
3	Destination Unreachable	13	Communication Administratively Prohibited
3	Destination Unreachable	14	Host Precedence Violation
3	Destination Unreachable	15	Precedence cutoff in effect
5	Redirect		

7	Unassigned		
8	Echo		
9	Router Advertisement		
10	Router Solicitation		
11	Time Exceeded		
12	Parameter Problem		
13	Timestamp		
14	Timestamp Reply		
40	Photuris	0	Bad SPI
40	Photuris	1	Authentication Failed
40	Photuris	2	Decompression Failed
40	Photuris	3	Decryption Failed
40	Photuris	4	Need Authentication
40	Photuris	5	Need Authorization
42-252	Unassigned		
255	Reserved		

The *Traceroute* command is used to discover the routes that packets take when traveling to their destination. In Taceroute the device sends out three UDP packets to an invalid port address at the remote host. Each packet has the Time-To-Live (TTL) field value set to one, which causes the datagram to "timeout" as soon as it hits the first router in the path. This router then responds with an ICMP Time Exceeded Message (Type 11 message) indicating that the datagram has expired. Another three UDP messages are now sent, each with the TTL value set to 2, which causes the second router to return ICMP Time Exceeded Messages. The process continues until the packets actually reach the other destination. Since these packets are trying to access an invalid port at the destination host, ICMP Port Unreachable Messages are returned, indicating an unreachable port; this event signals the Traceroute program that it is finished. The purpose behind this is to record the source of each ICMP Time Exceeded Message to provide a trace of the path the packet took to reach the destination. Here are some characters that may be output as feedback on a Traceroute command, all of which are based on ICMP messages.

Character	Description
nn msec	For each node, the round-trip time in milliseconds for the specified number of probes
*	The probe timed out
A	Administratively prohibited (example, access-list)
Q	Source quench (destination too busy)
I	User interrupted test
U	Port unreachable
H	Host unreachable
N	Network unreachable
P	Protocol Unreachable
T	Timeout
?	Unknown packet type

10. Conclusion

In this Chapter we have looked into the network layer and can see that this is a very important part of all networks. It involves every device in the network. We have learned the main functions of forwarding and routing, and the difference between the two, and discussed some principles of how some routing protocols work. We have learned about the two main network connection models – Virtual Circuit and Datagram networks – and how routing is performed within an AS and between ASs in Datagram networks (including the BGP). Also, it has been seen that the Internet requires the interworking of the Internet Protocol, the routing and control messaging to work. On the protocol side we have seen an overview of the structure of an IP datagram and the main distinguishing features between IPv4 and IPv6, as well as discussed what subnetting is. We have also discussed the purpose and working of

NAT, DHCP and ICMP. This now brings us further down the stack to the Link layer. But first, some review questions.

11. Review Questions

1. What is the difference between a segment, datagram, frame and packet?
2. What are the different services provided by the network layer to a flow of packets between a given source and destination?
3. How are Virtual Circuits implemented in a computer network?
4. What is a reason you can think of for why a packet cannot retain the same Virtual Circuit number on each of the links along its route?
5. How is a packet transmitted from source to destination in a datagram network?
6. What does each input port of a high speed router store to facilitate fast forwarding decisions?
7. What do you think are some advantages of switching via an interconnection network over switching via memory and switching via bus?
8. If a router has eight interfaces, how many IP addresses will it have? If more than one, what will the addresses have in common?
9. How can you determine whether a packet is being sent to a computer in the same network or to a computer in another network?
10. Suppose you purchase a wireless router and connect it to your ADSL connection. The ISP dynamically assigns your wireless router one IP address. If you have two PCs and three smartphones at home that use 802.11 (WiFi) to connect wirelessly to the router. How are IP addresses assigned to the PCs and phones? Does the wireless router use NAT? Why or why not?
11. Why are different inter-AS and intra-AS protocols used in the Internet?

Answers

1. The data unit used in the transport layer and passed up to the session or application layers, or passed down to the network layer is called a segment (e.g. a TCP segment). The network layer data unit is called a datagram, produced by

encapsulating a segment with network layer header. A frame is a data-link layer data unit, produced when stripping of the network layer header from a segment. A packet is a general term for a data unit. In TCP/IP a datagram sometime refers to a UDP data unit.

2. The network layer may provide to the higher layers an unreliable connectionless service where the sender and receiver treat each data unit independently through the datagram packet switching method, or a reliable connection-oriented service where the sender and receiver both see data as traveling on a logical connection and data is received in order, achieved through Virtual Circuit packet switching. Specific services may also include guaranteed delivery (with or without bounded delay), in-order packet delivery, guaranteed minimum bandwidth or maximum jitter and security services.
3. A VC requires a path from source to destination, implemented with routers and links mainly in the core of the network, VC numbers for each link along the path, and forwarding tables with entries for routers along the path. Whenever a new VC is established an entry is added to this forwarding table in the applicable routers.
4. By changing the number from link to link the length of the VC field is reduced since the number need not be carried. Also, by having a different VC number for each link along the path of a VC, network management is simplified. Each link in the path can choose a VC number independently of what the other links in the path chose. If a common number were required for all links along the path, the switches would have to exchange and process a substantial number of messages to agree on the VC number to be used for a connection.
5. Each packet is transmitted completely independently of other packets. Even if a packet is a part of multi-packet transmission the network treats it as though it existed alone. The source and destination address are used by the routers to decide the route for packets individually. All packets belonging to the same message may travel via different paths to reach the destination and so may arrive out of order and at different intervals
6. A shadow copy of the forwarding table computed by the routing processor is stored at each input port.

7. When switching by memory the speed is limited by the memory bandwidth. Similarly when switching over a bus the speed is limited by the bus bandwidth. Switching via interconnection network avoids these limitations
8. The router will have at least 8 IP addresses if the device capability allows this, which will have the same subnet with the first three octets of address being the same (assuming IPv4).
9. Look at the IP address and see if the first three octets are the same (assuming IPv4).
10. The router will probably have a DHCP server, which will dynamically assign a different IP address to each device. Yes, it must use NAT since the ISP only provides one IP address to the router.
11. It is mainly a matter of different goals and requirements for inter-AS and intra-AS communication. First, it is a matter of policy. Each AS is usually run by a different administrator and organisation. Among ASs it may be required that traffic from a specific AS not be able to pass into another, and control what traffic does pass through to another but within an AS this is not a concern. The scale of inter- and intra-AS routing is very different. Within an AS the scale is usually limited and not so much of a concern but scalability of protocols for inter-AS routing is critical.

12. Further reading

1. T. Pummill, Alantec, B. Manning, “Variable length subnet table for IPv4”, IETF, RFC 1878. December 1995. Accessible: https://tools.ietf.org/html/rfc1878
2. R. Gilligan, E. Nordmark, Sun Microsystems, Inc., “Transition Mechanisms for IPv6 Hosts and Routers”, IETF, RFC 1933, April 1996. Accessible: https://tools.ietf.org/html/rfc1933
3. R. Droms, Bucknell University, “Dynamic Host Configuration Protocol”, IETF, RFC2131, March 1997. Accessible: https://www.ietf.org/rfc/rfc2131.txt
4. J. Postel, ISI, “Internet Control Message Protocol”, IETF, RFC792, Accessible: https://tools.ietf.org/html/rfc792

5. F. Gont, UTN-FRH / SI6 Networks, C. Pignataro, Cisco Systems, “Formally Deprecating Some ICMPv4 Message Types”, IETF, RFC 6918. April 2013. Accessible: https://tools.ietf.org/html/rfc6918
6. J. Kurose and K. Ross, “Computer Networking: A Top-Down Approach (International) Chapter 3”, Pearson Education Ltd., Essex, sixth edition, 2013.