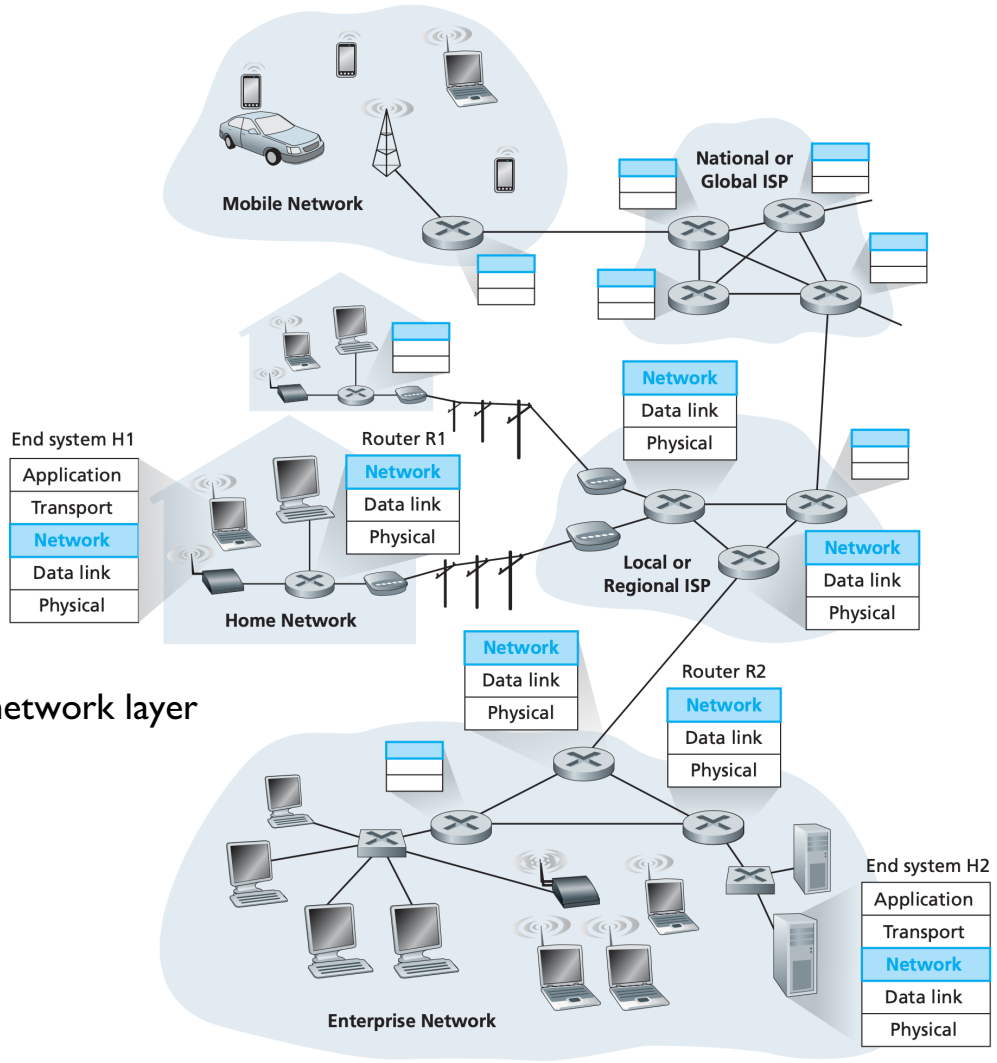
Network Layer

Hailiang Zhao @ ZJU.CS.CCNT
http://hliangzhao.me

This slide is a reminder for chapter 4 of the book *Computer Networking: A Top-Down Approach.*

Introduction

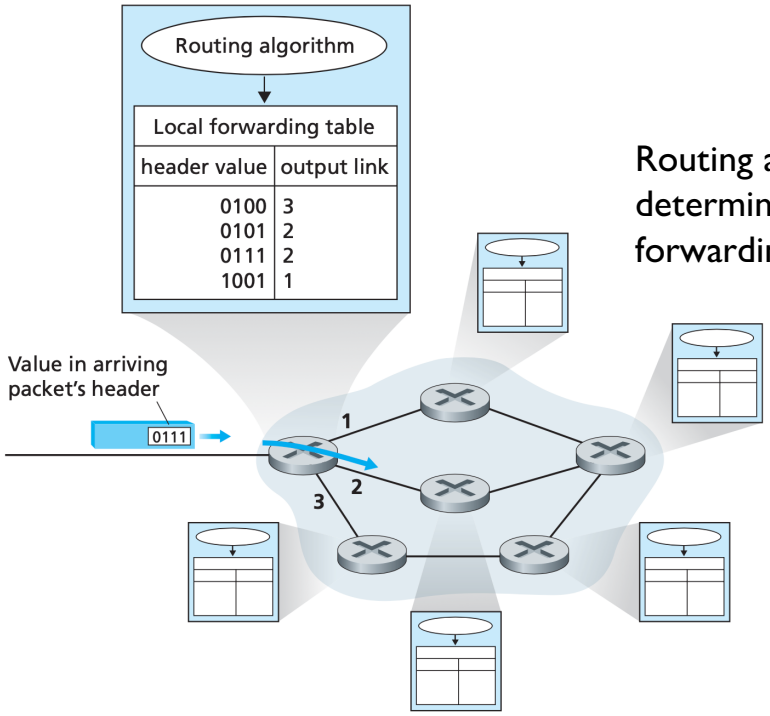
- We organize this chapter into three parts
  - network-layer functions and services (p. 3 ~ 7)
  - forwarding (p. 8 ~ ?)
  - routing (p.)



The network layer

Forwarding and Routing

- **Forwarding** refers to the router-local action of transferring a packet from an input link interface to the appropriate output link interface. **Routing** refers to the network-wide process that determines the end-to-end paths that packets take from source to destination
- Every router has a **forwarding table**. 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 forwarding table can configured by centralized or distributed routing algorithms (this is what routing protocol does)
- Distinguish *packet switch*, *link-layer switch*, and *router*
- Some network-layer architectures provide *connection setup* function



Routing algorithms determine values in forwarding tables

Network Service Models

- The Internet's network layer provides a single service, known as *best-effort service* ("no services at all")
- Other network architectures have defined and implemented service models that go beyond the Internet's best-effort service
  - the ATM network architecture [MFA Forum 2012, Black 1995] provides for multiple service models, meaning that different connections can be provided with different classes of service within the same network
    - Constant bit rate (CBR) ATM network service
    - Available bit rate (ABR) ATM network service

Network Architecture	Service Model	Bandwidth Guarantee	No-Loss Guarantee	Ordering	Timing	Congestion Indication
Internet	Best Effort	None	None	Any order possible	Not maintained	None
ATM	CBR	Guaranteed constant rate	Yes	In order	Maintained	Congestion will not occur
ATM	ABR	Guaranteed minimum	None	In order	Not maintained	Congestion indication provided

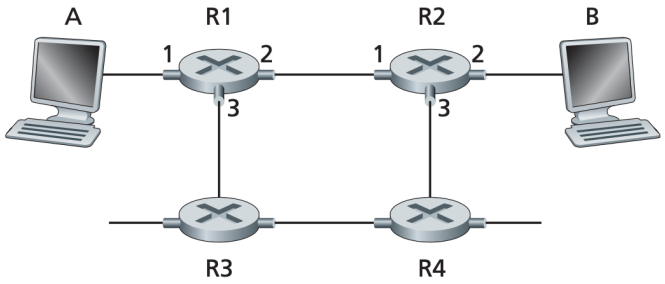
Internet, ATM CBR, and ATM ABR service models

Virtual Circuit and Datagram Networks

- A network layer can provide connectionless service or connection service between two hosts (not the same as transport-layer connection-oriented and connectionless services)
- Computer networks that provide only a connection service at the network layer are **called virtual-circuit (VC) networks**
- Computer networks that provide only a connectionless service at the network layer are called **datagram networks** [this is what Internet takes]
- VC network has its root in the telephony world
- Datagram network grew out of the need to connect computers together. The principle is simplifying the network infrastructure as far as possible because services are implemented at the network edge

Virtual Circuit and Datagram Networks

- Virtual-Circuit (VC) networks
  - A VC consists of (1) a path (that is, a series of links and routers) between the source and destination hosts, (2) VC numbers, one number for each link along the path, and (3) entries in the forwarding table in each router along the path
  - EXAMPLE:
    - Host A requests that the network establish a VC between itself and Host B
    - path: A-R1-R2-B
    - VC numbers: 12 22 32
    - In this case, when a packet in this VC leaves Host A, the value in the VC number field in the packet header is 12; when it leaves R1, the value is 22; and when it leaves R2, the value is 32

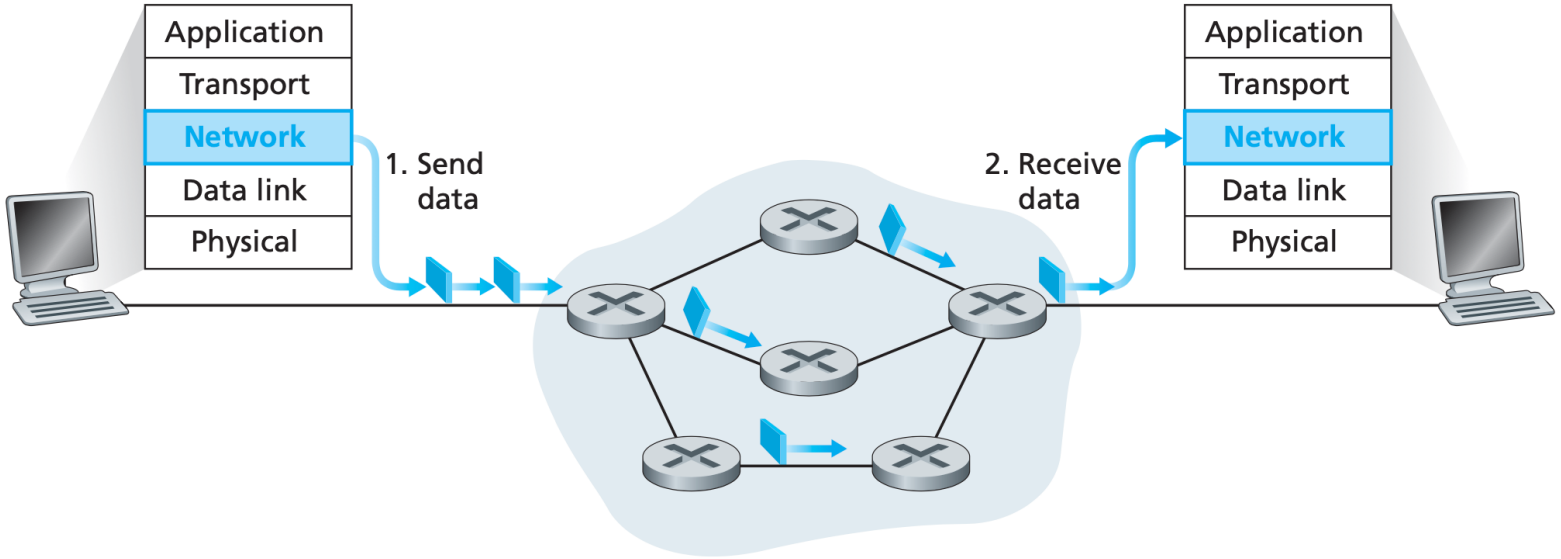


Incoming Interface	Incoming VC #	Outgoing Interface	Outgoing VC #
1	12	2	22
2	63	1	18
3	7	2	17
1	97	3	87
...	...	...	...

  - 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 across a router, a new connection entry must be added to the router's for- warding table; and each time a connection is released, an entry must be removed from the table

Virtual Circuit and Datagram Networks

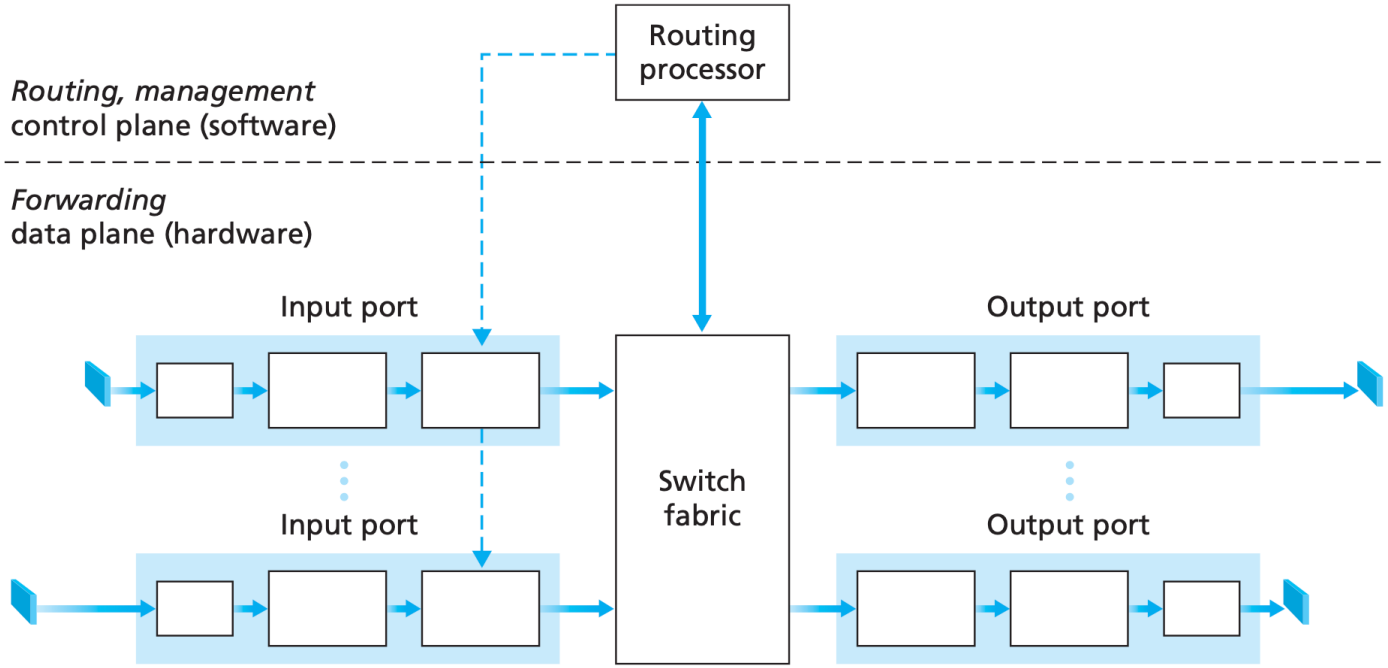
- Datagram networks
  - In a datagram network, each time an end system wants to send a packet, it stamps the packet with the address of the destination end system and then pops the packet into the network
  - Each router has a forwarding table that maps destination addresses to link interfaces; when 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
  - When there are multiple matches, the router uses the **longest prefix matching** rule



Datagram network

What is inside a Router?

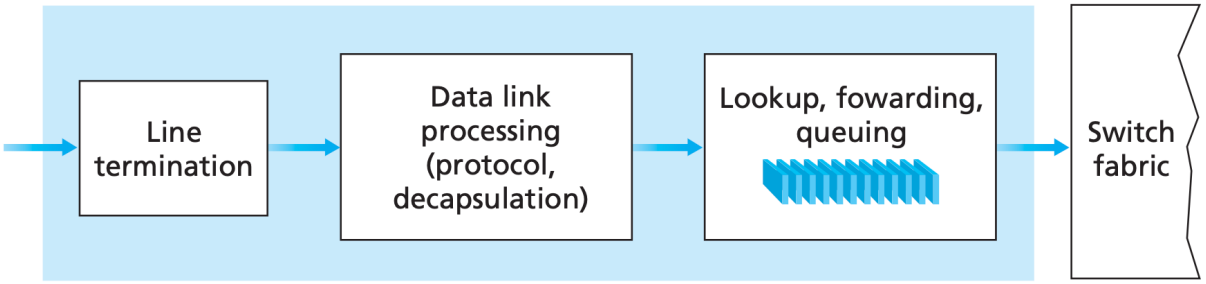
- Router forwarding plane
  - A router's input ports, output ports, and switching fabric together implement the forwarding function and are almost always implemented in hardware [scale of nanosecond]
- Router control plane
  - A router's control functions — executing the routing protocols, responding to attached links that go up or down, and performing management functions are implemented in software and execute on the routing processor (typically a traditional CPU) [scale of (milli)second]



Router architecture

Input Processing

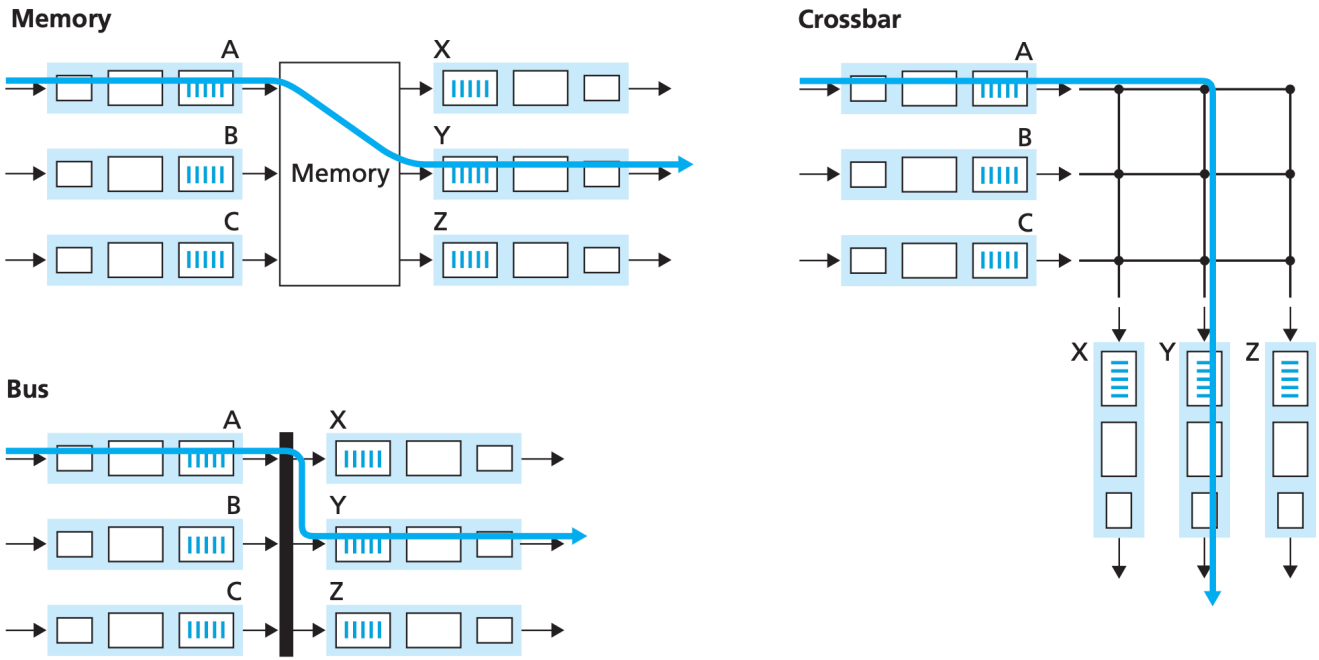
- The central to the router's operation: lookup
  - It is here that the router uses the forwarding table to look up the output port to which an arriving packet will be forwarded via the switching fabric
  - The forwarding table is computed and updated by the routing processor, with a shadow copy typically stored at each input port over a separate bus
  - Simply search through the forwarding table looking for the longest prefix match
  - The *Match-Plus-Action* abstraction
- Once a packet's output port has been determined via the lookup, the packet can be sent into the switching fabric
- A blocked packet will be queued at the input port and then scheduled to cross the fabric at a later point in time
- Beside lookup, many other actions are also taken: check and update the datagram header, etc.



Input port processing

Switching

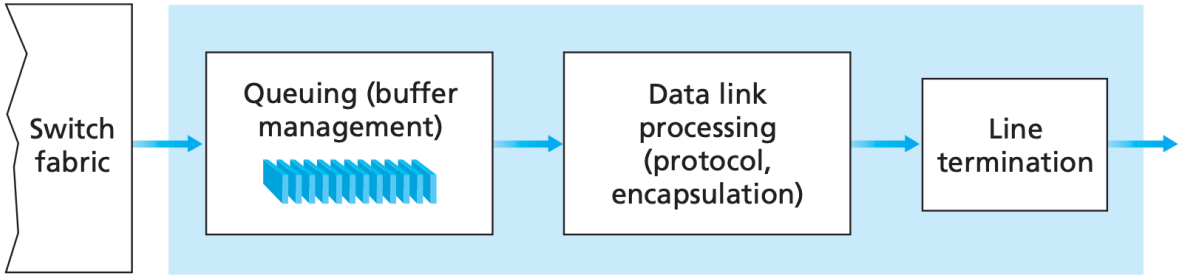
- Switch via memory
  - Input and output ports functioned as traditional I/O devices in a traditional operating system (signaled through interrupt)
  - Two packets cannot be forwarded at the same time
- Switch via a bus
  - If multiple packets arrive to the router at the same time, each at a different input port, all but one must wait since only one packet can cross the bus at a time
- Switch via an interconnection network
  - None-conflict switch can be done simultaneously



Three switching techniques

Output Processing

- Output port processing takes packets that have been stored in the output port's memory and transmits them over the output link
  - selecting and de-queueing packets for transmission
  - performing the needed link-layer and physical-layer transmission functions



Output port processing

Where Does Queuing Occur?

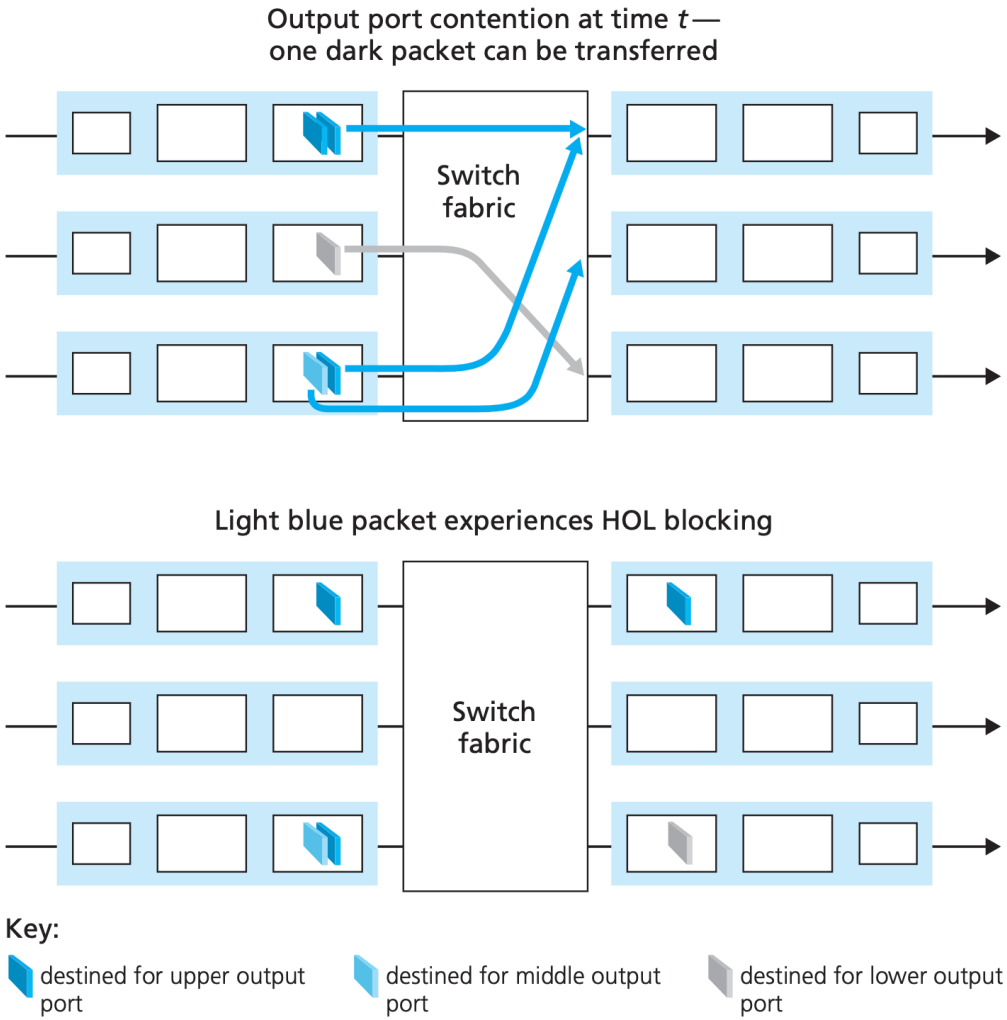
- The location and extent of queueing (either at the input port queues or the output port queues) will depend on the traffic load, the relative speed of the switching fabric, and the line speed
- For many years, the rule of thumb [RFC 3439] for buffer sizing was that the amount of buffering (B) should be equal to an average round-trip time (RTT, say 250 msec) times the link capacity (C)
- theoretical and experimental efforts [Appenzeller 2004], however, suggest that when
- there are a large number of TCP flows (N) passing through a link, the amount of buffering needed is

$$\frac{RTT \times C}{\sqrt{N}}$$

- Packet scheduler such as FCFS or weighted fair queuing (WFQ)
- Packet-dropping and -marking policies (aka active queue management) such as random early detection (RED)

Where Does Queuing Occur?

- **Head-of-line (HOL) Blocking**: a queued packet in an input queue must wait for transfer through the fabric (even though its output port is free) because it is blocked by another packet at the head of the line



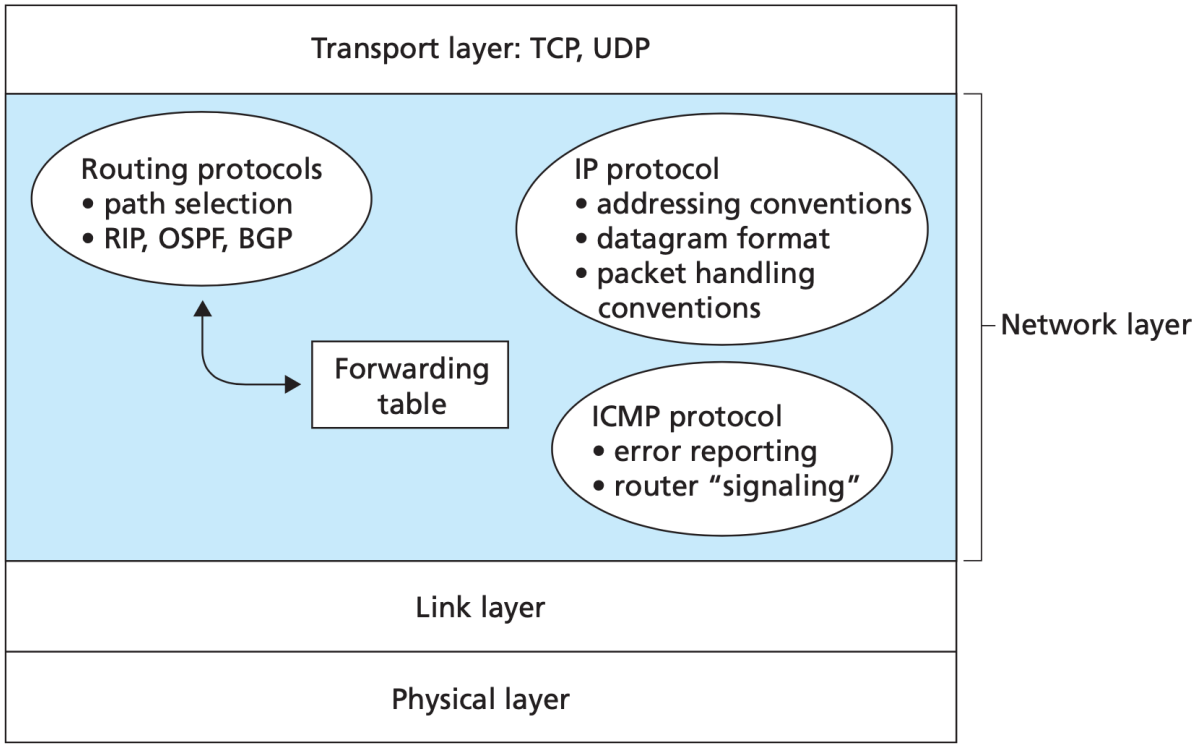
The Routing Control Plane

- Traditional routing control plane fully resides and executes in a routing processor within the router
- Recently, a number of researchers [Caesar 2005a, Casado 2009, McKeown 2008] have begun exploring new router control plane architectures
  - part of the control plane is implemented in the routers (e.g., local measurement/reporting of link state, forwarding table installation and maintenance) along with the data plane
  - part of the control plane can be implemented externally to the router (e.g., in a centralized server, which could perform route calculation)
  - A well-defined API dictates how these two parts interact and communicate with each other
  - Typical example: Software-Defined Network (SDN)

Forwarding and Addressing in the Internet

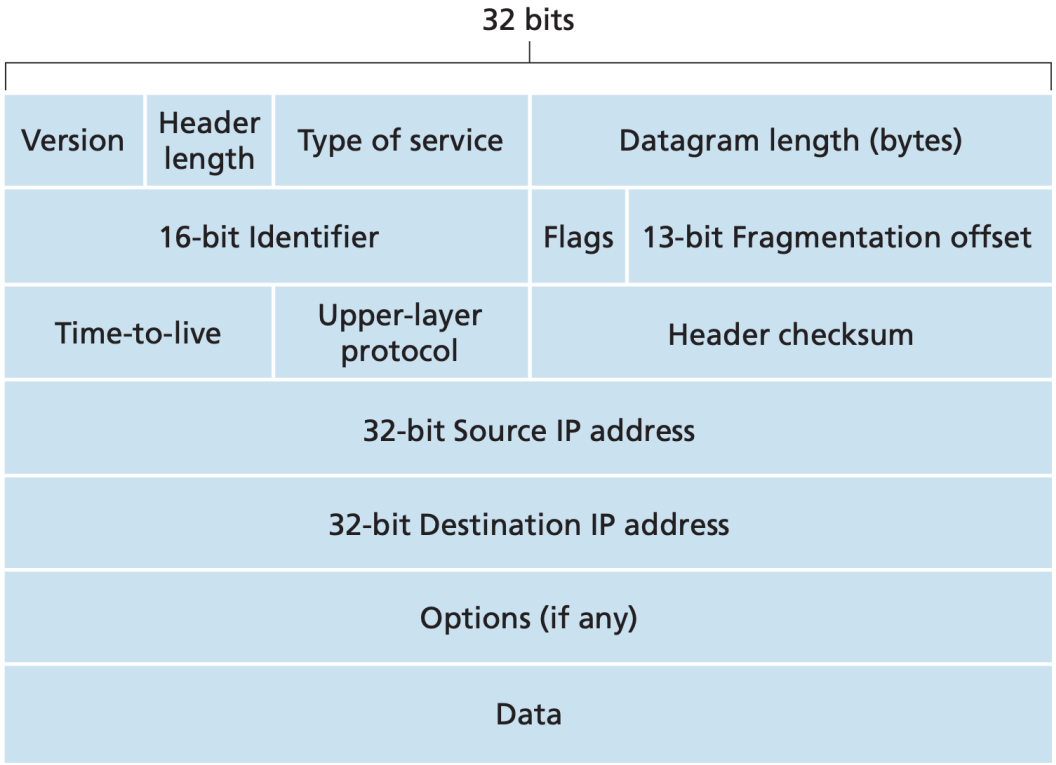
- Internet Protocol (IP)
  - addressing: IPv4, IPv6
  - packet handling
- Routing Protocols
  - determine the path a datagram follows from source to destination
- Internet Control Message Protocol (ICMP)
  - report errors in datagrams and respond to requests for certain network-layer information

A look inside the Internet's network layer



Datagram Format

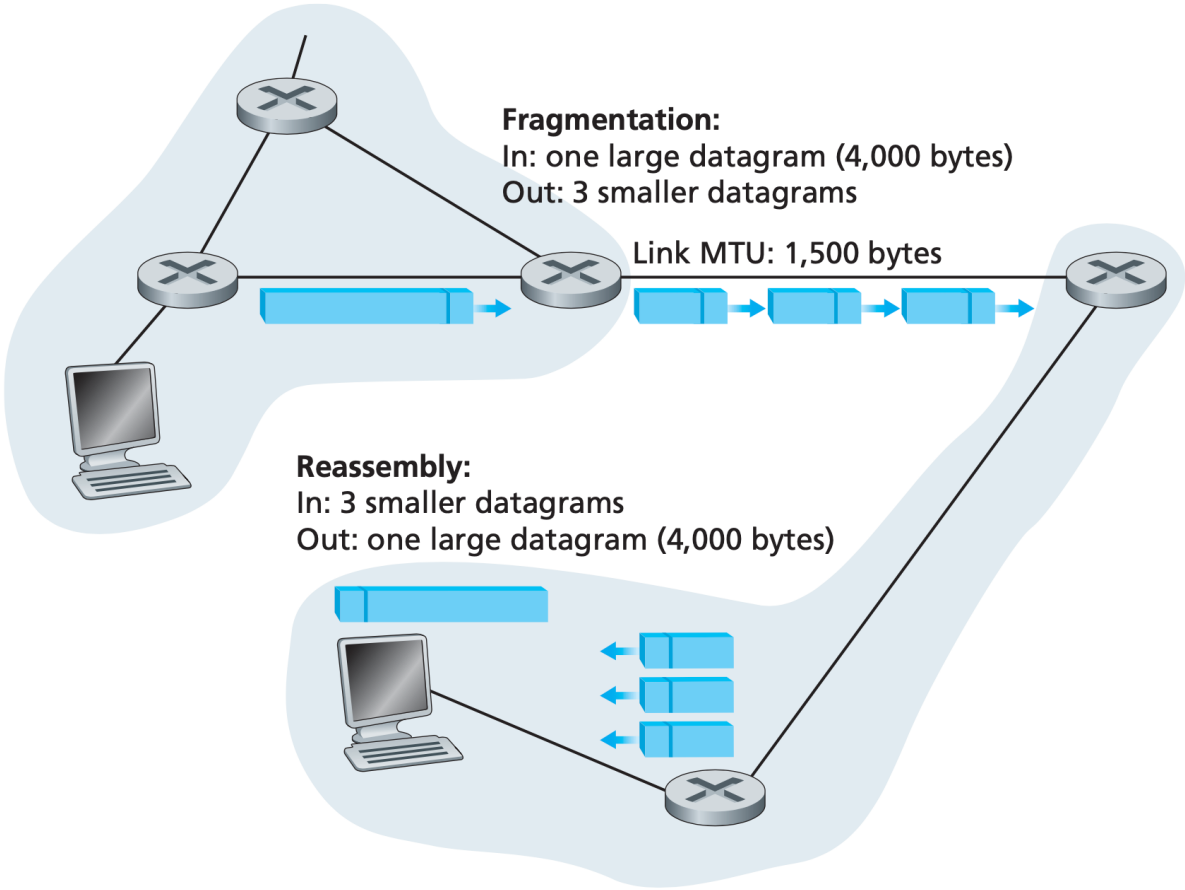
- Time-to-live: decremented by one each time the datagram is processed by a router
- Protocol: the glue that binds the network and transport layers together, whereas the port number is the glue that binds the transport and application layers together
- Data: transport-layer segment (TCP or UDP) or ICMP messages



IPv4 datagram format

IP Datagram Fragmentation

- IP fragmentation plays an important role in gluing together the many disparate link-layer technologies
- The maximum amount of data that a link-layer frame can carry is called the maximum transmission unit (MTU)



IP fragmentation and reassembly

IP Datagram Fragmentation

- IP fragmentation plays an important role in gluing together the many disparate link-layer technologies
- The maximum amount of data that a link-layer frame can carry is called the maximum transmission unit (MTU)

Fragment	Bytes	ID	Offset	Flag
1st fragment	1,480 bytes in the data field of the IP datagram	identification = 777	offset = 0 (meaning the data should be inserted beginning at byte 0)	flag = 1 (meaning there is more)
2nd fragment	1,480 bytes of data	identification = 777	offset = 185 (meaning the data should be inserted beginning at byte 1,480. Note that 185 · 8 = 1,480)	flag = 1 (meaning there is more)
3rd fragment	1,020 bytes (= 3,980–1,480–1,480) of data	identification = 777	offset = 370 (meaning the data should be inserted beginning at byte 2,960. Note that 370 · 8 = 2,960)	flag = 0 (meaning this is the last fragment)

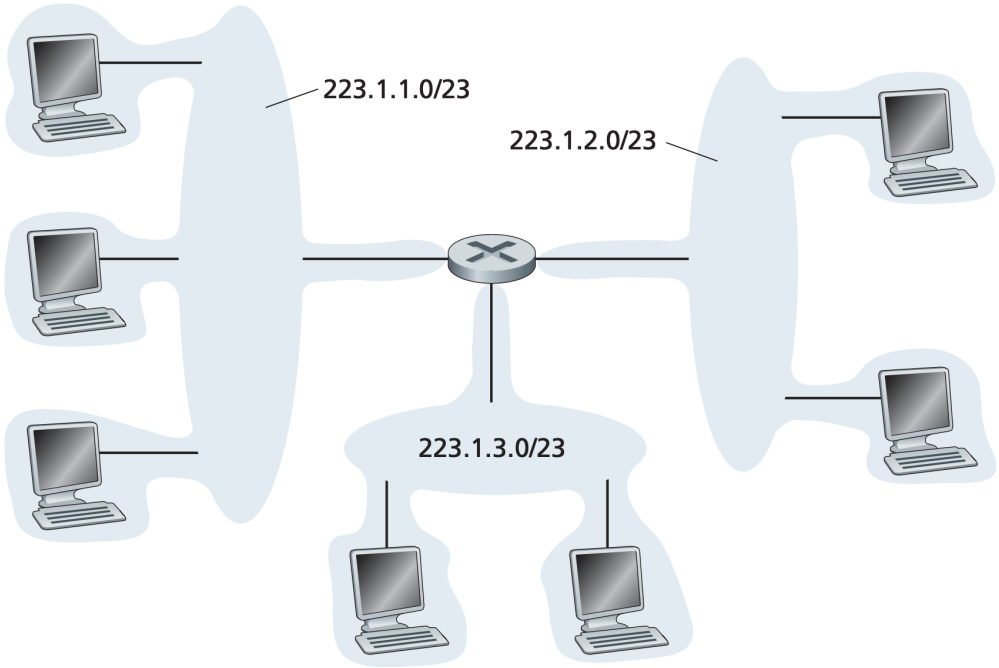
IP fragments

IPv4 Addressing

- The boundary between the router and any one of its links is also called an *interface*; The boundary between the host and the physical link is also called an *interface*
- An IP address is technically **associated with an interface**, rather than with the host or router containing that interface
- IP address is written in **dotted-decimal notation**
- Each interface on every host and router in the global Internet must have an IP address that is globally unique (except for interfaces behind NATs)
- A portion of an interface's IP address will be determined by the subnet to which it is connected

IPv4 Addressing

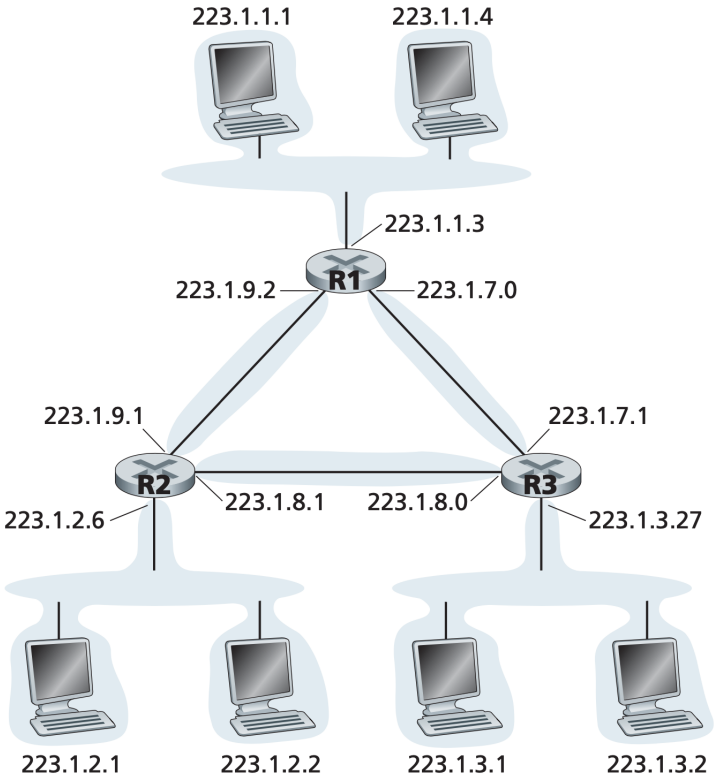
- The fig. below provides an example of IP addressing and interfaces
- The four interfaces having the address of the form of 223.1.1.x forms a *subnet* (aka a *IP network* or a *network*). They are also interconnected to each other by a network that contains no routers
- The most popular interconnection tech is **Ethernet LAN**, in which case the interfaces would be interconnected by an Ethernet switch, or by a wireless access point
- IP addressing assigns an address to this subnet: 223.1.1.0/24, where the /24 notation, sometimes known as a **subnet mask**



Subnet addressing

IPv4 Addressing

- The IP definition of a subnet is not restricted to Ethernet segments that connect multiple hosts to a router interface
- The fig. below has 6 subnets. The subnet 223.1.9.0/24 which interconnects R1 and R2 is not using Ethernet LAN
- *To determine the subnets, detach each interface from its host or router, creating islands of isolated networks, with interfaces terminating the end points of the isolated networks. Each of these isolated networks is called a subnet.*



Three routers interconnecting six subnets

IPv4 Addressing

- How addressing is handled?
  - The Internet's address assignment strategy is known as Classless Interdomain Routing (CIDR—pronounced cider) [RFC 4632]
  - The x most significant bits of an address of the form a.b.c.d/x constitute the network portion of the IP address, and are often referred to as the prefix (or network prefix) of the address
  - An organization is typically assigned a block of contiguous addresses, that is, a range of addresses with a common prefix
  - When a router outside the organization forwards a datagram whose destination address is inside the organization, only the leading x bits of the address need be considered. This considerably reduces the size of the forwarding table in these routers
  - The remaining 32-x bits of an address can be thought of as distinguishing among the devices within the organization, all of which have the same network prefix
  - These lower-order bits may (or may not) have an additional subnetting structure
  - Before CIDR was adopted, the network portions of an IP address were constrained to be 8, 16, or 24 bits in length, an addressing scheme known as **classful addressing**, since subnets with 8-, 16-, and 24-bit subnet addresses were known as class A, B, and C networks, respectively [C too small, B too large]
  - When a host sends a datagram with destination address 255.255.255.255, the message is delivered to all hosts on the same subnet [the IP broadcast address]

IPv4 Addressing

- How addressing is handled? (cont'd)
  - Obtaining a Block of Addresses for an Organization
    - The network administrator might first contact its ISP, which would provide addresses from a larger block of addresses that had already been allocated to the ISP
    - As shown below, The ISP, in turn, could divide its address block into eight equal-sized contiguous address blocks and give one of these address blocks out to each of up to eight organizations
    - IP addresses are managed under the authority of the Internet Corporation for Assigned Names and Numbers (ICANN)
    - The role of the nonprofit ICANN organization is not only to allocate IP addresses, but also to manage the DNS root servers. It also has the very contentious job of assigning domain names and resolving domain name disputes

ISP's block	200.23.16.0/20	11001000 00010111 0001**0000** 00000000
Organization 0	200.23.16.0/23	11001000 00010111 0001000**0** 00000000
Organization 1	200.23.18.0/23	11001000 00010111 0001001**0** 00000000
Organization 2	200.23.20.0/23	11001000 00010111 0001010**0** 00000000
. . .	. . .	. . .
Organization 7	200.23.30.0/23	11001000 00010111 0001111**0** 00000000

Dividing into eight equal-sized contiguous address blocks

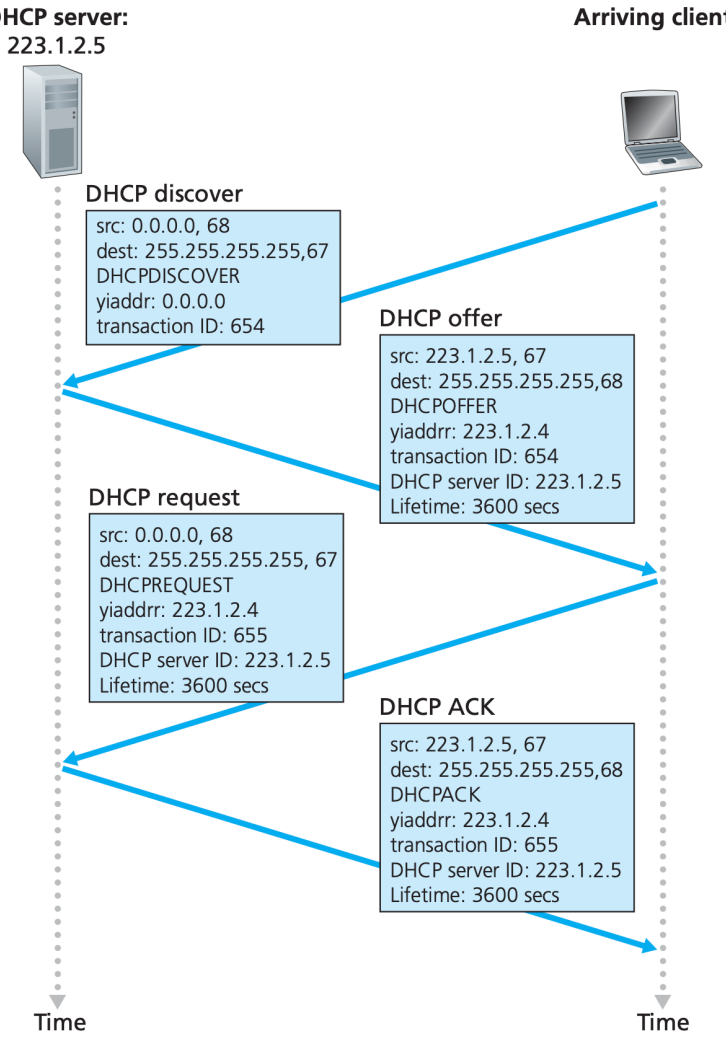
IPv4 Addressing

- How addressing is handled? (cont'd)
  - Obtaining a Host Address: the Dynamic Host Configuration Protocol
    - A system administrator will typically manually configure the IP addresses with the organization into the router (often remotely, with a network management tool)
    - Host addresses can also be configured manually, but more often this task is now done using the Dynamic Host Configuration Protocol (DHCP)
    - A network administrator can configure DHCP so that a given host receives the same IP address each time it connects to the network, or a host may be assigned a temporary IP address that will be different each time the host connects to the network
    - DHCP also allows a host to learn additional information, such as its subnet mask, the address of its first-hop router (often called the default gateway), and the address of its local DNS server
    - As a plug-and-play protocol, DHCP is also enjoying widespread use in residential Internet access networks and in wireless LANs

IPv4 Addressing

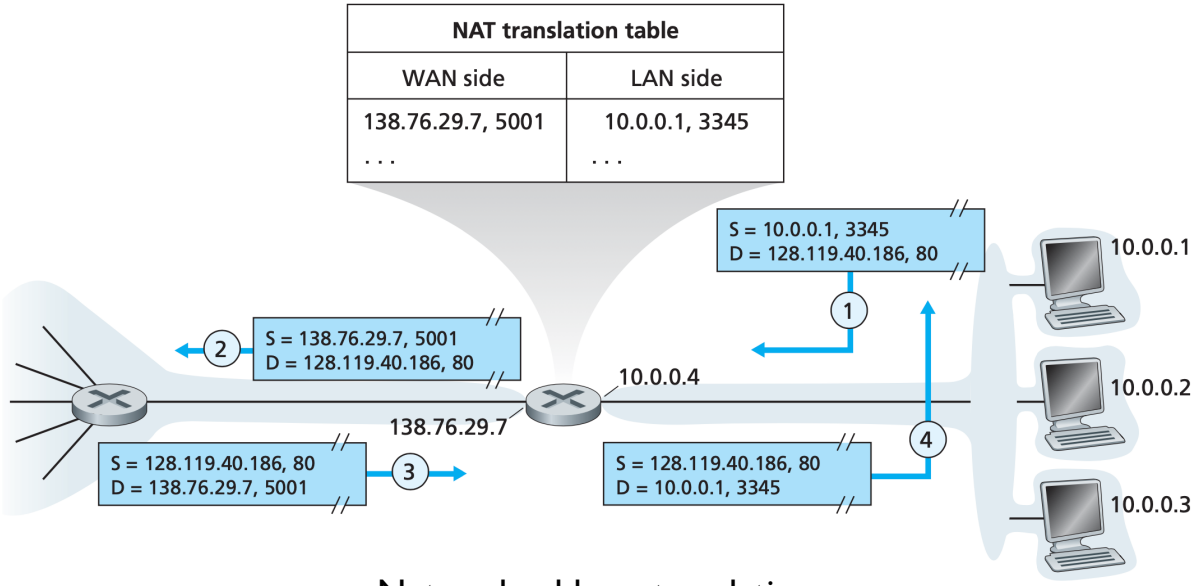
- How addressing is handled? (cont'd)
  - Obtaining a Host Address: the Dynamic Host Configuration Protocol (cont'd)
    - DHCP is a client-server protocol
    - Each subnet will have a DHCP server or a DHCP relay agent (typically a router)
    - The four-step process:
    - Connection will be break when moving between diff. subnets

DHCP client-server interaction



IPv4 Addressing

- Network Address Translation (NAT)
  - The address space 10.0.0.0/8 is one of three portions of the IP address space that is reserved in [RFC 1918] for a private network or a realm with private addresses
  - Devices within a given home network can send packets to each other using 10.0.0.0/24 addressing. However, packets forwarded beyond the home network into the larger global Internet clearly cannot use these addresses (as either a source or a destination address) because there are hundreds of thousands of networks using this block of addresses (only meaningful within the subnet)
  - The NAT-enabled router does not look like a router to the outside world. Instead the NAT router behaves to the outside world as a single device with a single IP address. The router gets its address from the ISP's DHCP server
  - NAT translation table is used to distinguish internal hosts



Network address translation

IPv4 Addressing

- Network Address Translation (NAT) (cont'd)
  - Many purists in the IETF community loudly object to NAT
    - Port numbers are meant to be used for addressing processes, not for addressing hosts
    - the NAT protocol violates the so-called end-to-end argument; that is, hosts should be talking directly with each other, without interfering nodes modifying IP addresses and port numbers
    - routers are supposed to process packets only up to layer 3
    - we should use IPv6 to solve the shortage of IP addresses, rather than recklessly patching up the problem with a stopgap solution like NAT
    - **Hosts behind a NAT cannot act as a server and accept TCP connections** (interferes with P2P applications)
  - UPnP
    - NAT traversal is increasingly provided by Universal Plug and Play (UPnP), which is a protocol that allows a host to discover and configure a nearby NAT
    - UPnP allows external hosts to initiate communication sessions to NATed hosts, using either TCP or UDP

Internet Control Message Protocol (ICMP)

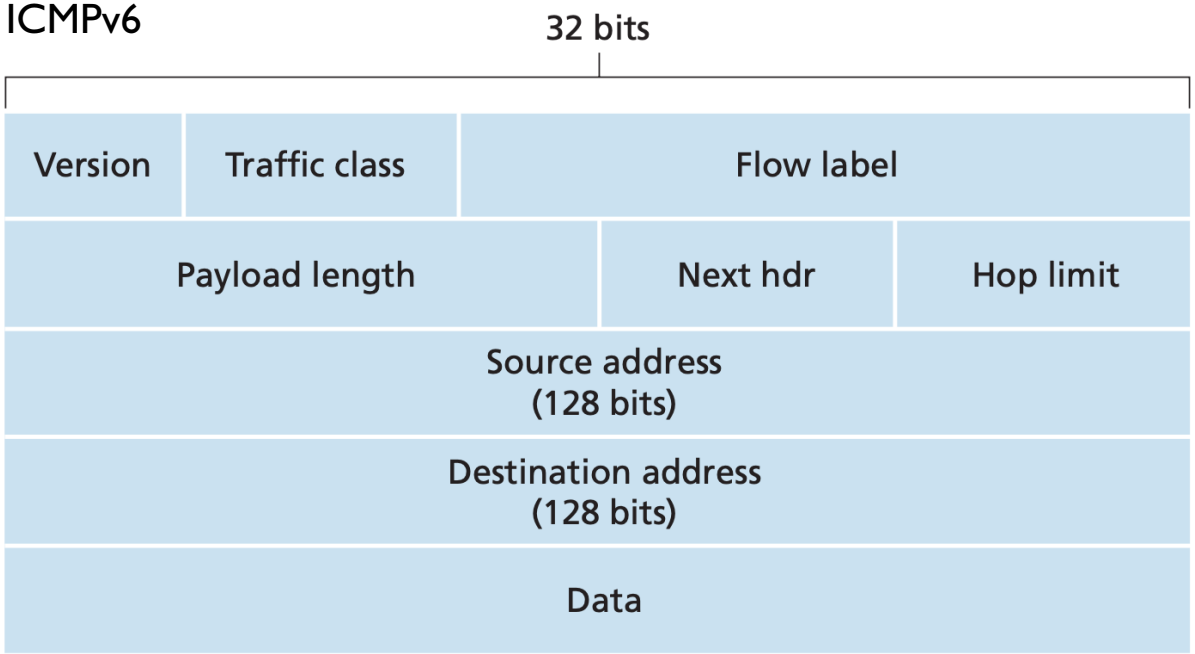
- ICMP, specified in [RFC 792], is used by hosts and routers to communicate network-layer information to each other (error reporting, source quench msg)
- ICMP messages are carried as IP payload, just as TCP or UDP segments are carried as IP payload [considered part of IP but architecturally lies above IP]
- ICMP messages have a type and a code field, and contain the header and the first 8 bytes of the IP datagram that caused the ICMP message to be generated in the first place
- Traceroute and Ping are implemented with ICMP messages

ICMP Type	Code	Description
0	0	echo reply (to ping)
3	0	destination network unreachable
3	1	destination host unreachable
3	2	destination protocol unreachable
3	3	destination port unreachable
3	6	destination network unknown
3	7	destination host unknown
4	0	source quench (congestion control)
8	0	echo request
9	0	router advertisement
10	0	router discovery
11	0	TTL expired
12	0	IP header bad

ICMP message types

IPv6 Datagram Format

- The most important changes introduced in IPv6
  - Expanded addressing capabilities
  - A streamlined fixed-length, 40-byte header
  - Flow labeling and priority
- Fields appearing in the IPv4 datagram but are no longer present in the IPv6 datagram
  - Fragmentation/Reassembly: Replaced by "Packet Too Big" ICMP error message back to the sender
  - Header checksum: recompute of checksum because of the TTL filed is time-consuming
  - Options
- ICMP → ICMPv6



IPv6 datagram format