Network layer

Datagram is a network layer packet

Network layer does forwarding and routing

When a packet arrives at a router's input link, the router must move the packet to the appropriate output link (forwarding)

The network layer must determine the route or path taken by packets as they flow from a sender to a receiver(Routing)

____

Every router has a forwarding table(?)
How to forward:

○ Examine value of a field in the arriving packet's header. Depending on protocol, this can be either the destination or indication of connection to which packet belongs.

○ Use this value to index into router's forwarding table(FT)

The value in FT has the outgoing link interface(i.e. through which link to send it)

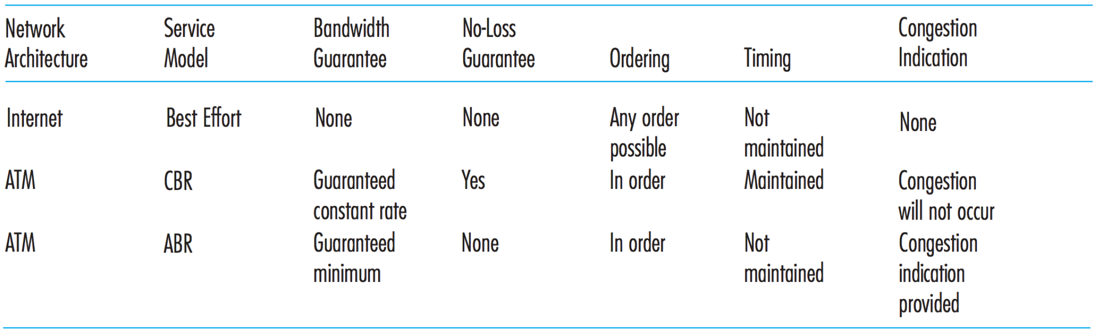
____

How are FT calculated?
Routing algorithm(RA) does it.
RA are centralised(whole RA runs on a central site) or decentralised(some parts run on each of the routers)
Router gets **routing protocol messages** which are used to configure FT.

Another important work that Network layer does is connection setup.
What is connection setup?
Some network-layer architectures - like ATM, frame relay, and MPLS - 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-to-destination connection can begin to flow. In the network layer, this process is referred to as connection setup.

Guarantees that network layer can provide:

○ Guaranteed delivery
○ Guaranteed delivery(bounded time)
○ In-order packet delivery(FIFO)
○ Minimal bandwidth(if transmission is done at lower bitrate than Minimal bandwidth, no packet loss)
○ Max jitter(Amount of time between 2 transmissions is same as amount of time between 2 receptions)
○ Security(Session keys)



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

Constant bitrate(CBR), Available bitrate(ABR)

CBR used by telephone companies.

○ Network layer services are host-to-host, transport layer’s are process-to-process.

○ Either connectionless(datagram networks) or connection service(Virtual circuit network), but not both.

○ The network-layer connection service is implemented in the routers in the net- work core as well as in the end systems.

Internet is datagram network while ATM and frame relay are virtual circuits(VC).

____

A VC consists of:

○ A path(links and routers) between source and destination.

○ VC numbers(1 # for each link along paths)

○ Entries in forwarding table at each router(Each row is a 4-tuple i.e. Incoming interface, Incoming VC #, outgoing interface, outgoing VC #)

A packet should have VC number. Each router checks Incoming interface and VC #, sees table and determines outgoing interface and replaces VC # with the outgoing one.

Whenever a new VC is established across a router, an entry is added to the forward- ing table. Similarly, whenever a VC terminates, the appropriate entries in each table along its path are removed.

Why does a packet not keep the same VC number on each of the links along its route?

○ Replacing the number each time makes VC # field shorted and saves space

○ This way, all routers don’t have to agree on common VC #s and saves time for new VC setup

3 phases in VC:

○ VC setup

○ Data transfer

○ VC teardown

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 **signaling messages** and the protocols used to exchange these messages are often referred to as **signaling protocols.**

In VC, connection state information is stored.

____

In Datagram, FT has destination address -> link interface
For 32 bit addresses(Like IPv4) each router needs to store 2**32 - 1 (> 4e9)
So, we store prefixes of common addresses in the table, such that the longest prefix match can determine which interface to go to. In this case, if there are k interfaces, only k entries are needed in the table.
In Datagram, forwarding state info is stored.
In a datagram network the forwarding tables are modified by the routing algorithms, which typically update a forwarding table every 1-5 mins or so.

____

4 Router components:

○ Input ports - performs physical layer function of terminating an incoming physical link at a router, also performs link-layer functions needed to interoperate with the link layer, also does lookup in forwarding table. Control packets are forwarded from from input port to routing processor

○ Switching fabric - connects input ports to output ones

○ Output ports - If bidirectional links, these are paired with input port

- ○ Routing processor - Executes routing protocol, maintains routing tables and attached link state information, and computes the forwarding table

These forwarding functions are sometimes collectively referred to as the **router forwarding plane**.

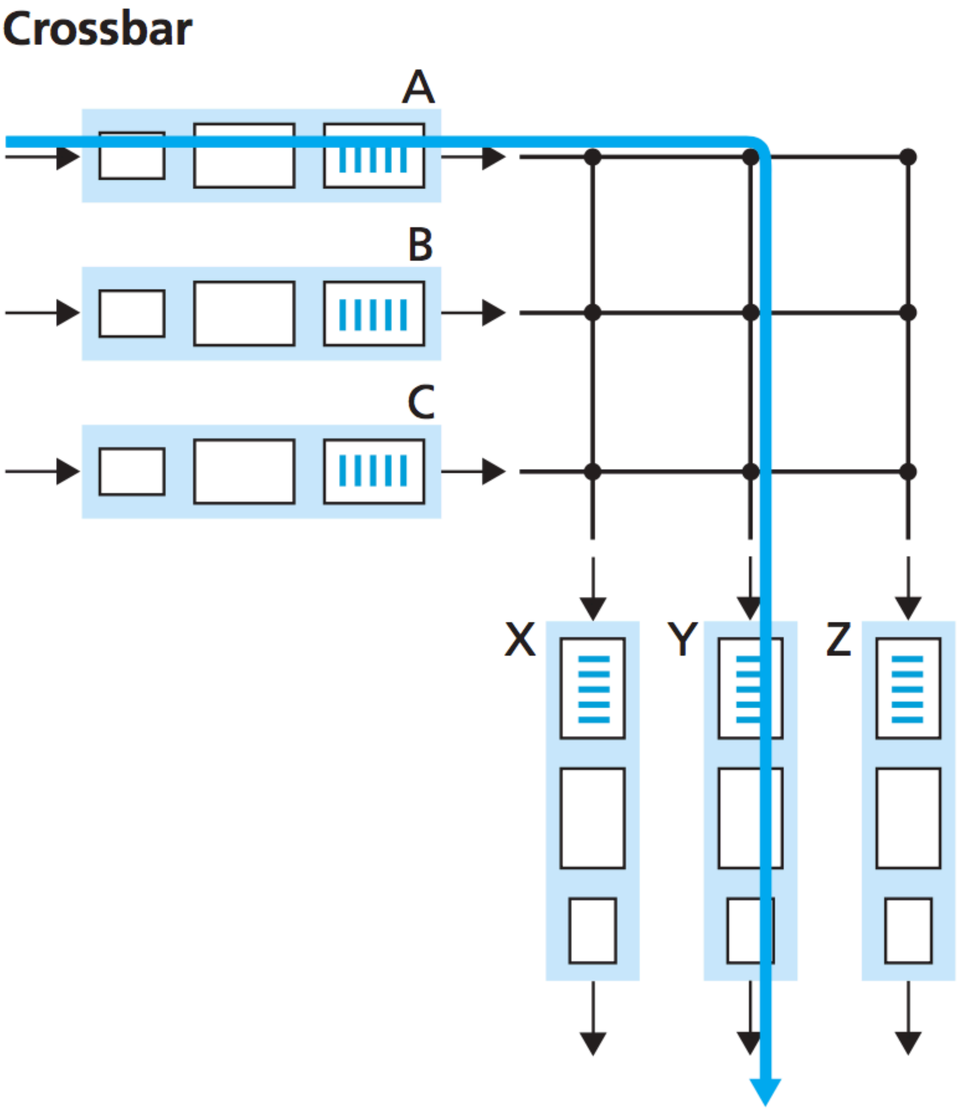
Special attention must also be paid to memory access times, resulting in designs with embedded on-chip DRAM and faster SRAM (used as a DRAM cache) memories. Ternary Content Address Memories (TCAMs) are also often used for lookup.
With a TCAM, a 32-bit IP address is presented to the memory, which returns the content of the forwarding table entry for that address in essentially constant time. The Cisco 8500 has a 64K CAM for each input port.

How is switching done?

- ○ Switching via memory - earlier routers were PCs so switching was done via CPUs. Input and output ports functioned as traditional I/O devices in a traditional operating system. In this scenario, if the memory bandwidth is such that *B* packets per second can be written into, or read from, memory, then the overall forwarding throughput (the total rate at which packets are transferred from input ports to output ports) must be less than *B*/2. Note also that two packets cannot be forwarded at the same time, even if they have different destination ports, since only one memory read/write over the shared system bus can be done at a time.
- ○ Switching via bus - an input port transfers a packet directly to the output port over a shared 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. Because every packet must cross the single bus, the switching speed of the router is limited

to the bus speed.

○ Switching via interconnection network - To overcome bandwidth limitation on Bus. A crossbar switch is an interconnection net- work consisting of 2*N* buses that connect *N* input ports to *N* output ports.



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 *B = RTT C/*sqrt(N). N is the min # of input/output ports available.

A consequence of output port queuing is that a **packet scheduler** at the output port must choose one packet among those queued for transmission.

○ Can be FCFS or weighted fair queuing (WFQ), which shares the outgoing link fairly among the different end-to-end connections that have packets queued for transmission. Packet scheduling plays a crucial role in providing **quality-of-service guarantees**

In some cases, it may be advantageous to drop (or mark the header of) a packet *before* the buffer is full in order to provide a congestion signal to the sender.

____

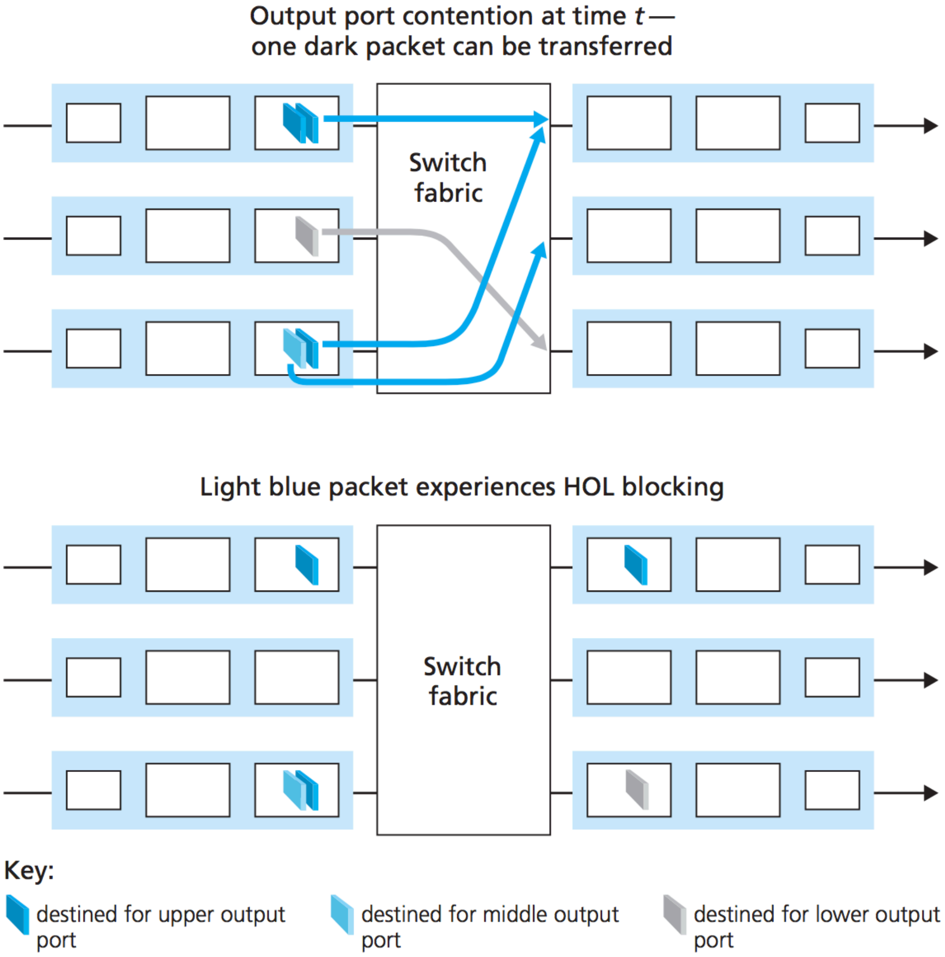
AQM(**active queue management)** algorithms is the **Random Early Detection** (**RED**) algorithm

A weighted average is maintained for the length of the output queue.

○ If the average queue length is less than a minimum threshold*, min*, when a packet arrives, the packet is admitted to the queue

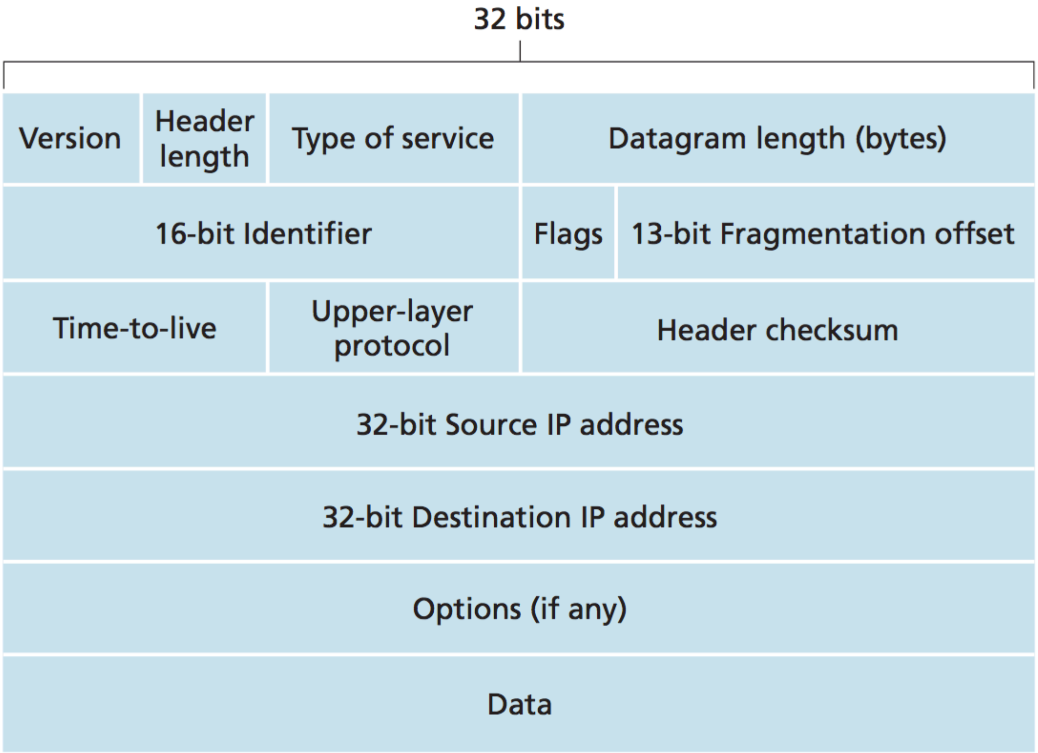
○ If the packet arrives to find an average queue length in the interval [*min*, *max*], the packet is marked or dropped with a probability that is typically some function of the average queue length, *min*, and *max*

Head of line (HOL) blocking


Output port contention at time t—
one dark packet can be transferred
Switch fabric
Light blue packet experiences HOL blocking
Switch fabric
Key:
destined for upper output port
destined for middle output port
destined for lower output port

Due to HOL blocking, the input queue will grow to unbounded length under certain assumptions as soon as the packet arrival rate on the input links reaches only 58 percent of their capacity.

____
____



IPv4 datagram

- ○ Version: 4 bits. Specify the IP protocol version.
- ○ Header length: Because an IPv4 datagram can contain a variable number of options (which are included in the IPv4 datagram header), these 4 bits are needed to determine where in the IP datagram the data actually begins. Typically 20 bytes of header
- ○ Type of service: The type of service (TOS) bits were included in the IPv4 header to allow different types of IP datagrams to be distinguished from each other. 8 bits
- ○ Datagram length: This is the total length of the IP datagram (header plus data), measured in bytes. 16 bits.
- ○ Identifier(16 bits), flags(3 bits), fragmentation offset(13): These three fields have to do with so-called IP fragmentation.
- ○ Time-to-live: The time-to-live (TTL) field is included to ensure that datagrams do not circulate forever. 8 bits.
- ○ Protocol: This field is used only when an IP datagram reaches its final destination. Which transport layer to use.

TCP - 6 and UDP - 17. Analogous to port #

○ Header checksum: each 2 bytes. Routers discard if errors. Note that only the IP header is checksummed at the IP layer, while the TCP/UDP checksum is computed over the entire TCP/UDP segment. Second, TCP/UDP and IP do not necessarily both have to belong to the same protocol stack.

○ Options: The options fields allow an IP header to be extended.

○ Data - payload.

20 bytes of header(just as in TCP) without options.

Ethernet frames can carry up to 1,500 bytes of data.
The maximum amount of data that a link-layer frame can carry is called the maximum transmission unit (MTU).

Sticking to the principle of keeping the network core simple, the designers of IPv4 decided to put the job of datagram reassembly in the end systems rather than in network routers.

When a datagram is created, the sending host stamps the datagram with an identification number as well as source and destination addresses. Typically, the sending host increments the identification number for each datagram it sends.
The last fragment has a flag bit set to 0, whereas all the other fragments have this flag bit set to 1.

In order for the destination host to determine whether a fragment is missing (and also to be able to reassemble the fragments in their proper order), the offset field is used to specify where the fragment fits within the original IP datagram.
Offset is multiplied by 8 to get byte number, so only fragment at multiples of 8

Fragmentation can be used to create lethal DoS attacks, whereby the attacker sends a series of bizarre and unexpected fragments. A classic example is the Jolt2 attack, where the attacker sends a stream of small fragments to the target host, none of which has an offset of zero. The target can collapse as it attempts to rebuild datagrams out of the degenerate packets. Another class of exploits sends overlapping IP fragments, that is, fragments whose offset values are set so that the fragments do not align properly.

____

The boundary between the host and the physical link is called an **interface**.
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)

In IP terms, this network interconnecting three host interfaces and one router interface forms a **subnet.** IP addressing assigns an address to this subnet: 223.1.1.0/24, where the /24 notation, sometimes known as a **subnet mask**, indicates that the leftmost 24 bits of the 32-bit quantity define the subnet address.

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**.

The Internet's address assignment strategy is known as **Classless Interdomain Routing** (**CIDR**—pronounced *cider*). Before CIDR was adopted, the network portions of an IP address were con- strained 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.

With subnet of size X, one can buy 2**x - 2 (2 are reserved for special purposes).

Internet Corporation for Assigned Names and Numbers (ICANN) manages IP addresses.

____

Host addresses can also be configured manually, but more often this task is now done using the **Dynamic Host Configuration Protocol (DHCP).**
In addition to host IP address assignment, 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.
DHCP is a **plug-and-play protocol**

DHCP is a client-server protocol. A client is typically a newly arriving host wanting to obtain network configuration information, including an IP address for itself.

DHCP has 4 steps:

○ DHCP server discovery: Find DHCP server. Client sends UDP package to port 67. Destination IP is 255.255.255.255 and source IP is 0.0.0.0. This is sent to all hosts by the client. Called **DHCP discover message**

○ DHCP server offer: Sends **DHCP offer message.** Each server 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.

○ DHCP request: The newly arriving client will choose from among one or more server offers and respond to its selected offer with a **DHCP request message**, echoing back the configuration parameters.

○ DHCP ACK: The server responds to the DHCP request message with a **DHCP ACK message**, confirming the requested parameters

____

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.
NAT uses DHCP as well.

Against NAT:

○ Port numbers are meant to be used for addressing processes, not for addressing hosts.

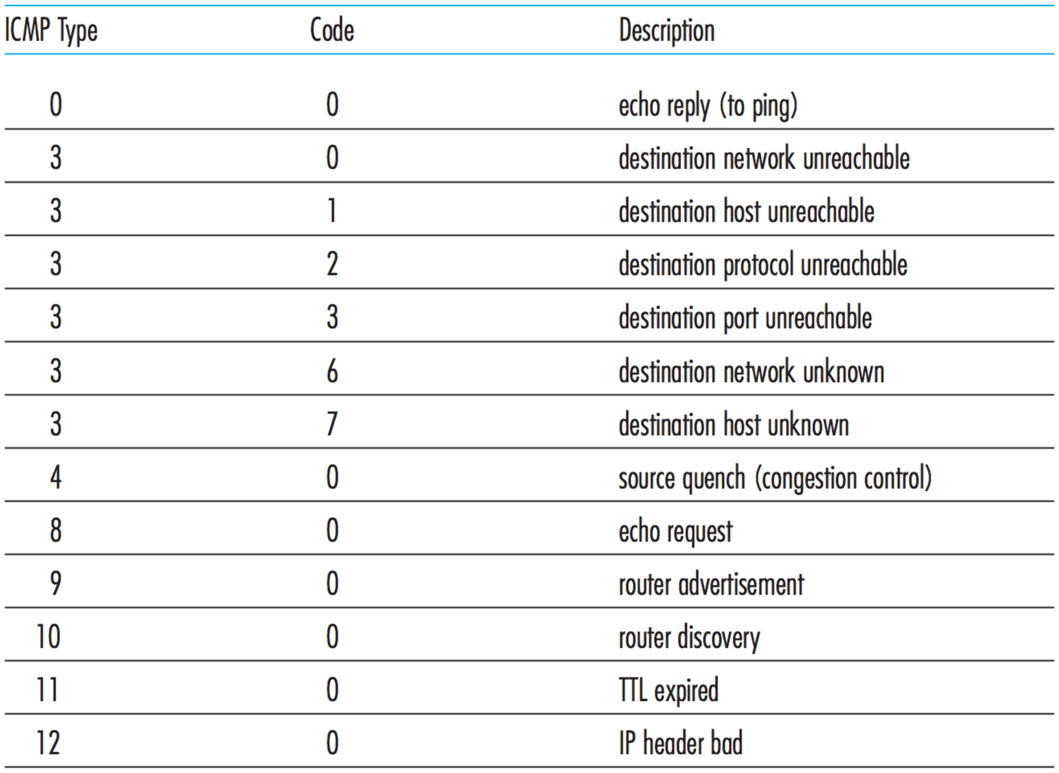
○ Routers are supposed to process packets only up to layer 3

○ Third, they argue, the NAT protocol violates the so-called end-to-end argument; that is, hosts should be talk- ing directly with each other, without interfering nodes modifying IP addresses and port numbers.

○ And fourth, they argue, we should use IPv6 instead

____

Type 3 icmp means error
ICMP messages are carried as IP payload
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
The well-known ping program sends an ICMP type 8 code 0 message to the specified host. The destination host, seeing the echo request, sends back a type 0 code 0 ICMP echo reply
ICMP source quench message. Original purpose was to perform congestion control.
Allow a congested router to send an ICMP source quench message to a host to force that host to reduce its transmission rate.

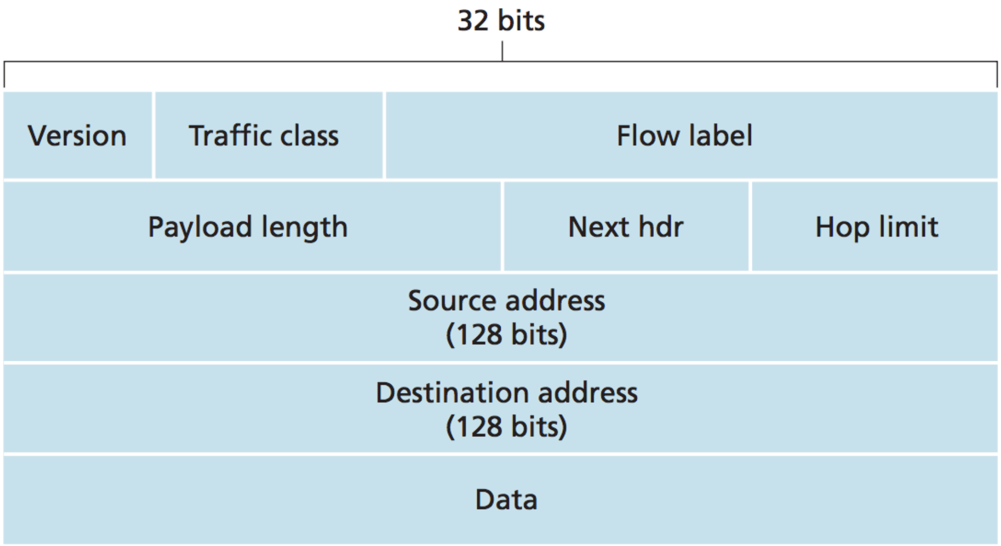


ICMP Type	Code	Description
0	0	echo reply (to ping)
3	0	destination network unreachable
3	1	destination host unreachable
3	2	destination protocol unreachable
3	3	destination port unreachable
3	6	destination network unknown
3	7	destination host unknown
4	0	source quench (congestion control)
8	0	echo request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

How does traceroute stop?
Recall that the source increments the TTL field for each datagram it sends. Thus, one of the datagrams will eventually make it all the way to the destination host. Because this datagram contains a UDP segment with an unlikely port number, the destination host sends a port unreachable ICMP message (type 3 code 3) back to the source.

____

IPv6 has 128 bits
Unicast, multicast and anycast address
A streamlined 40-byte header



Traffic class. This 8-bit field is similar in spirit to the TOS field we saw in IPv4.

Flow label. As discussed above, this 20-bit field is used to identify a flow of
datagrams

Payload length. This 16-bit value is treated as an unsigned integer giving the number of bytes in the IPv6 datagram following the fixed-length, 40-byte data- gram header.

Next header. UDP or TCP. Same as IPv4

Hop limit.

New version of ICMP - IGMP Internet Group Management Protocol