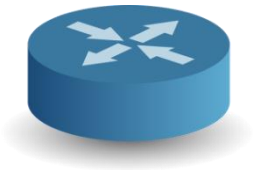
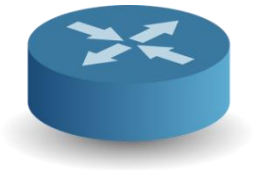


Computer Networks

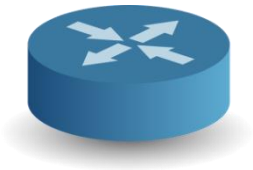
Part Three

Network Layer in the Internet



Network Layer in the Internet

- IP Version 4
- IP Addresses
- IP Version 6
- Internet Control Protocols
- Label Switching and MPLS



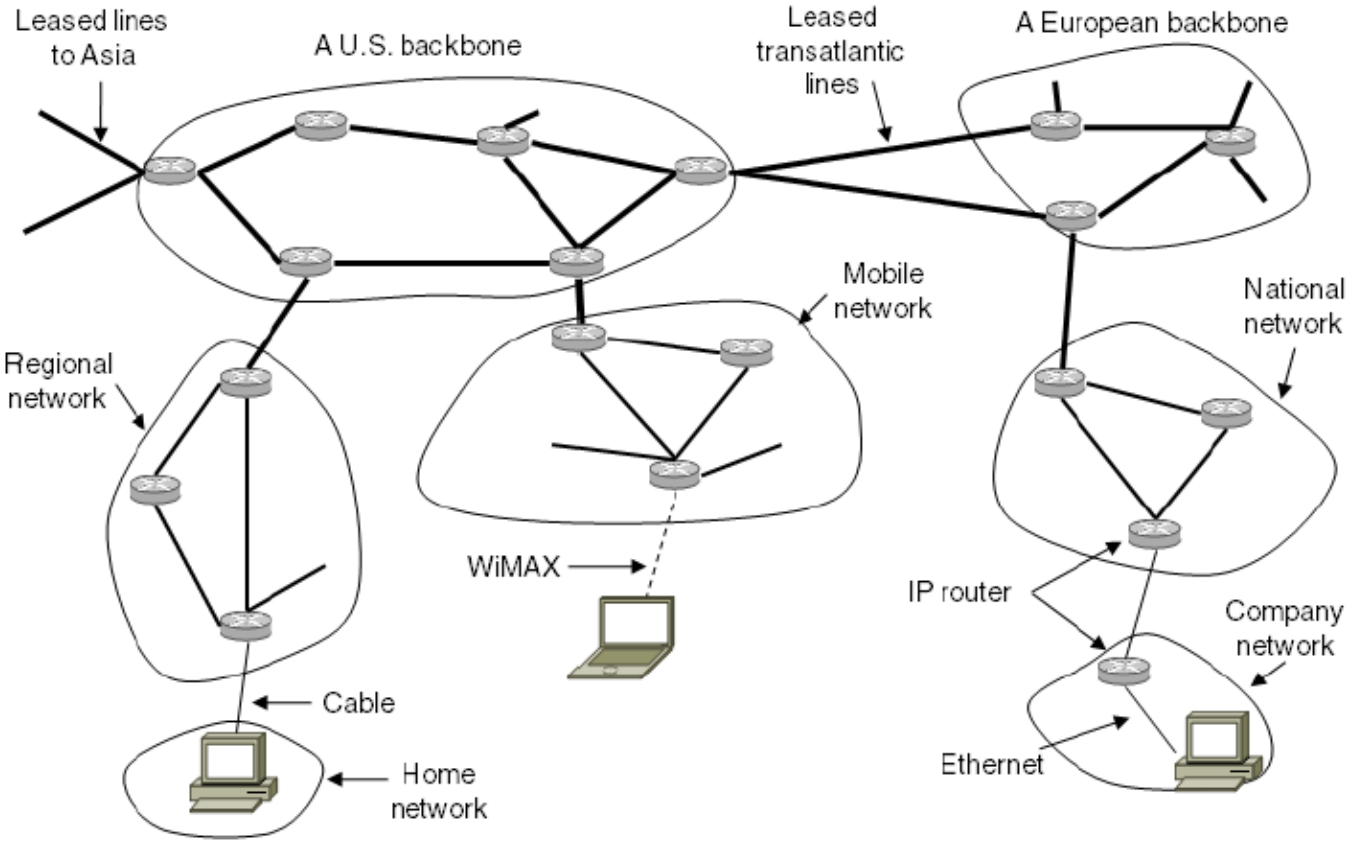
Network Layer in the Internet

IP has been shaped by guiding principles:

- Make sure it works
- Keep it simple
- Make clear choices
- Exploit modularity
- Expect heterogeneity
- Avoid static options and parameters
- Look for good design (not perfect)
- Strict sending, tolerant receiving
- Think about scalability
- Consider performance and cost

Network Layer in the Internet (3)

Internet is an interconnected collection of many networks that is held together by the IP protocol



IP Datagram Layout

a) Header contains
   - Source Internet address
   - Destination Internet address
   - Datagram type field
   - Other useful information

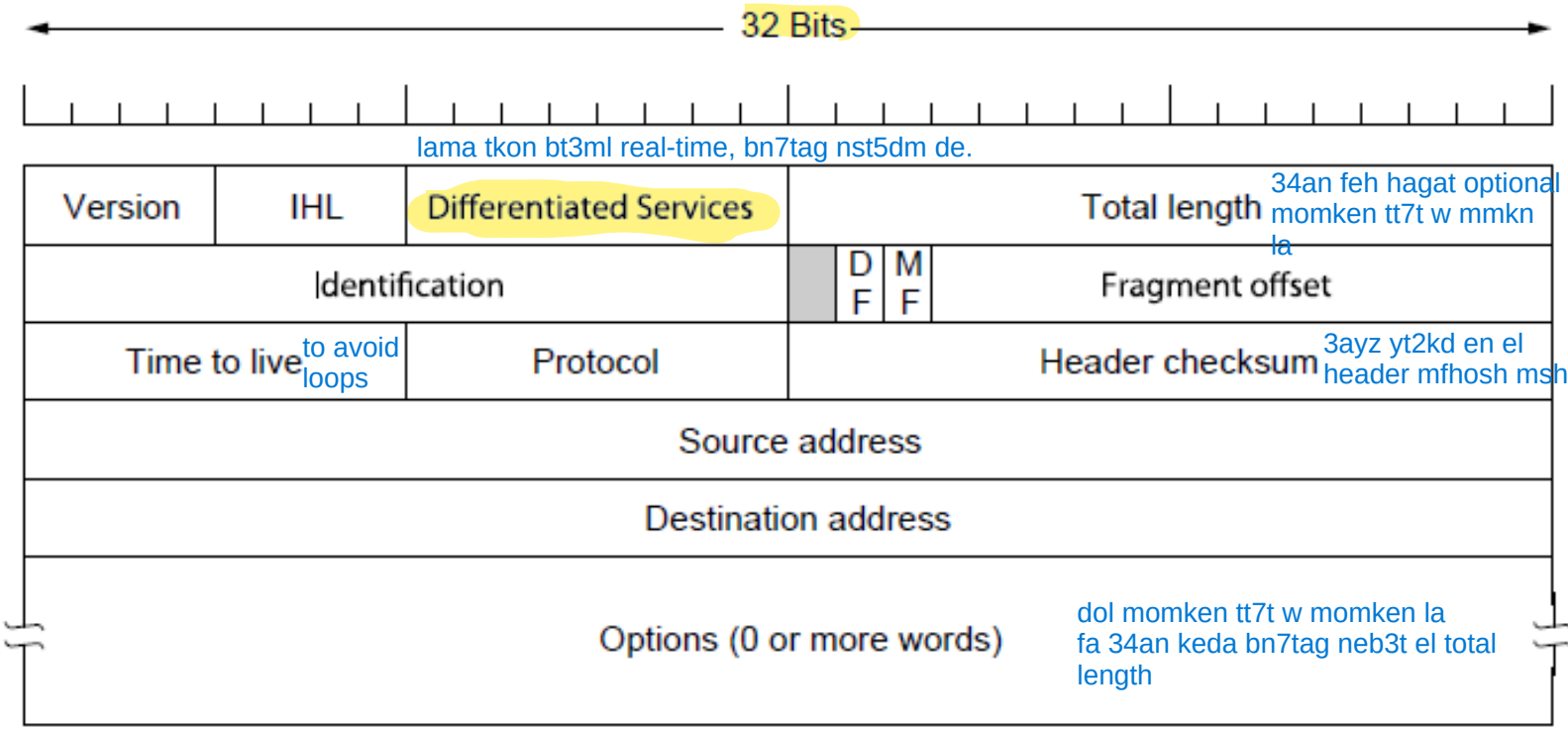
b) Payload contains data being carried

c) IP is a **best effort** protocol

IP Version 4 Protocol

IPv4 (Internet Protocol) header is carried on all packets and has fields for the key parts of the protocol:

4 bytes.



-----
total 5 rows = 20 bytes

fragment y3ny by2sm el data.
fa 34an ye2dr ygm3hom, lazm ykon mdehom id.

yeb2a hwa bysm7 bl fragmantation, by2sm el data w yb3tha, w 3nd el reciever, aana m7tag agm3hom bl id, w artbhom bl offset, w 34an a3rf en lsa feh packets, bla2y el more fragment (MF) b 1, lw msh 3auz a2at3ha hab3t flag dont fragment (DF).

IP Fragmentation (RFC791)

e7na bnlg2 lel fragmantation lama ykon el allowed lengh lel packet fe autonomous system ana ray7lo a2al mn el allowed length 3ndy.

- If the DF bit is set and the packet needs to be fragmented, the packet is dropped
- All fragments have **MF** bit set **except the last fragment**
- All fragments belonging to the same original packet have the same value in the "*identification*" field
- The **first fragmenting node** sets the "*identification*" field
- Each fragment is a totally independent datagram
  - Must contain all information needed to deliver the packet to the destination
  - Must contain the original protocol value
  - The header must be calculated correctly
  - Fragmenting router may override/modify the **TypeOfService** TOS bits (Differentiated Service)
  - Options from the original datagram may be copied to the fragments
- Fragments of the same datagram are uniquely identified by
  - Identification, Source address,
  - Destination address
  - protocol

IP Fragmentation (RFC791)

- "Fragment Offset" contains the offset of the first byte of the payload of a fragment from the beginning of the original packet
  - In multiple of **8 octets**
- A fragment can be *further fragmented* along the way
  - This means that MF bit must have the **SAME VALUE** as the larger datagram except for the very last fragment
  - The identification field remains the same
- Reassembly occurs at the destination
  - The destination must allocate buffer equal to the maximum datagram size
  - First fragment has offset zero
  - Last fragment has MF bit cleared
  - Each fragment has the offset, so it can be placed correctly

Fragmentation

- RULE :
  - The amount of data sent in one fragment is chosen such that It is as large as possible but less than or equal to MTU including 20 bytes header
  - It is a multiple of 8 so that pure decimal value can be obtained for the fragment offset field

Fragmentation Example

Consider There is a host A present in network X having MTU = 520 bytes. There is a host B present in network Y having MTU = 200 bytes.
Host A wants to send a message to host B.
Consider router receives a datagram from host A having-

Header length = 20 bytes, Payload length = 500 bytes, Total length = 520 bytes, DF bit set to 0

Sol: Router decides to send maximum 176 bytes of data in one fragment instead of 180 bytes (multiple of 8) +20 B header

- Fragment 1 : 176B, Fragment offset field value = 0, MF= 1
- Fragment 2: 176B, Fragment offset field value = 176 / 8 = 22, MF= 1
- Fragment 3: 148B, Fragment offset field value = (176 + 176) / 8 = 44, MF= 0

The IP Protocol (IP options) (RFC791)

Complete list of options can be found in

www.iana.org/assignments/ip-parameters

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

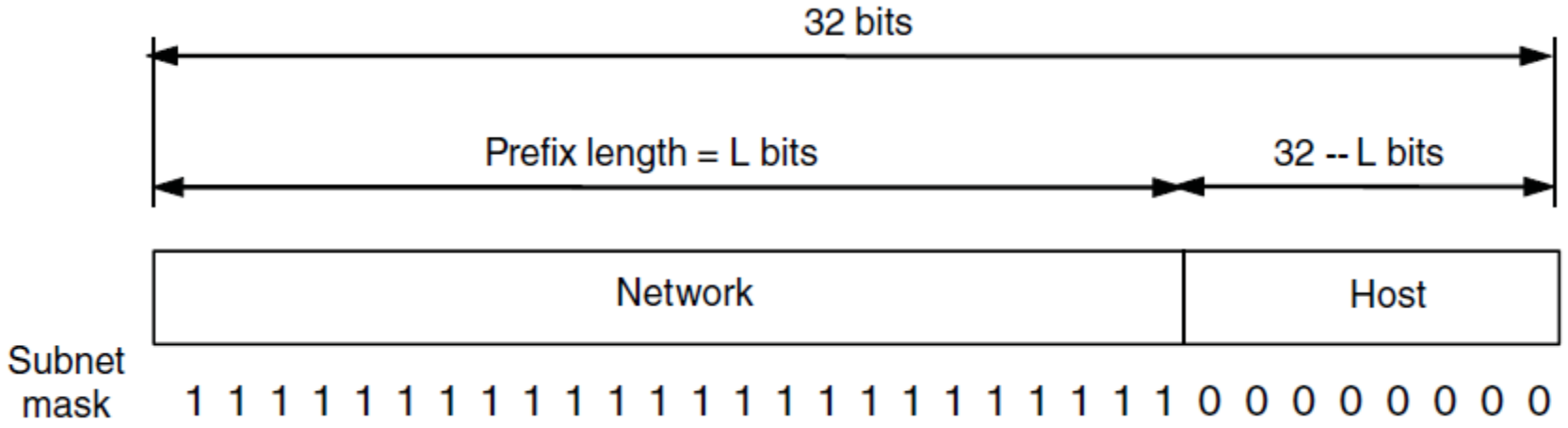
a) Options may be formatted as
   - Single octet
   - TLV: type, length, value
     - Type (1 octet): option type
     - Length (1 octet): total length of the option, including the type field
     - Value: the actual bytes of the option
     - Type contains a “copied” bit. Set when the option is copied to all fragments

b) More details in the RFC

IP Addresses – Prefixes

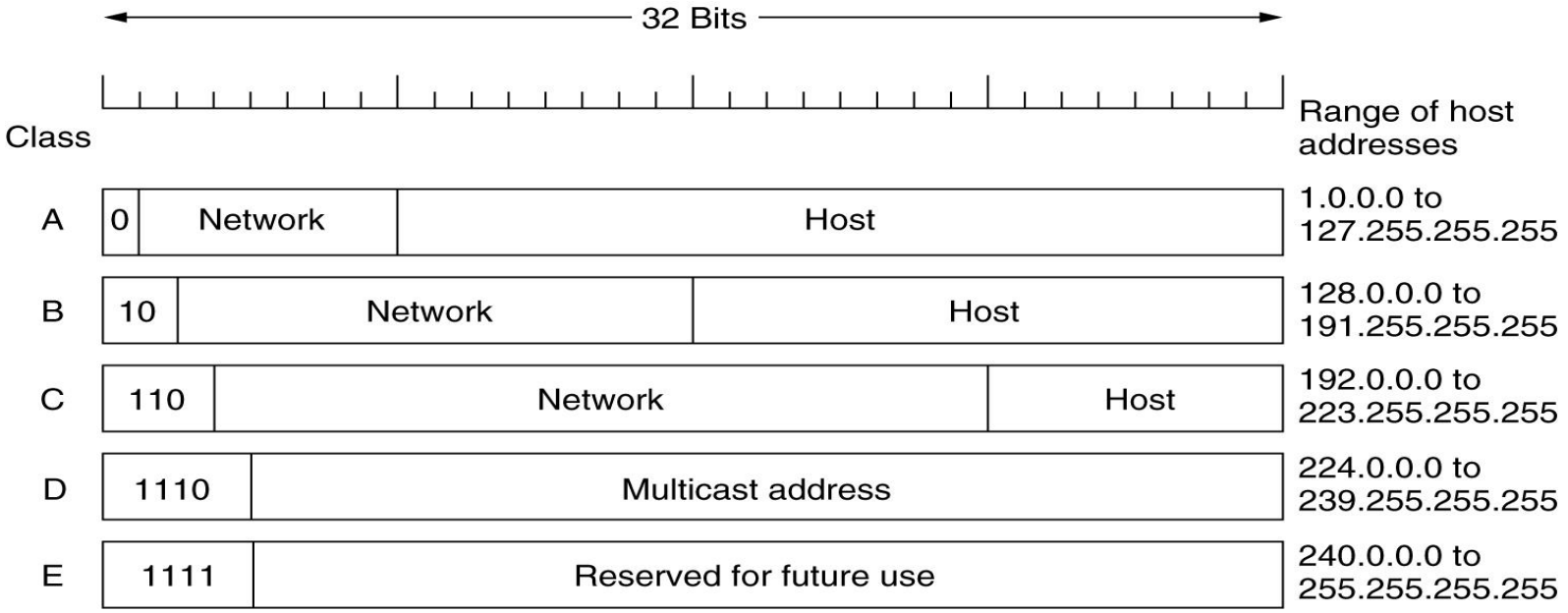
Addresses are allocated in blocks called prefixes

- Prefix is determined by the network portion
- Has 2^L addresses aligned on 2^L boundary
- Written address/length, e.g., 18.0.31.0/24 (3 bytes address)



IPv4 Addresses: Classful Addressing

IP address formats.



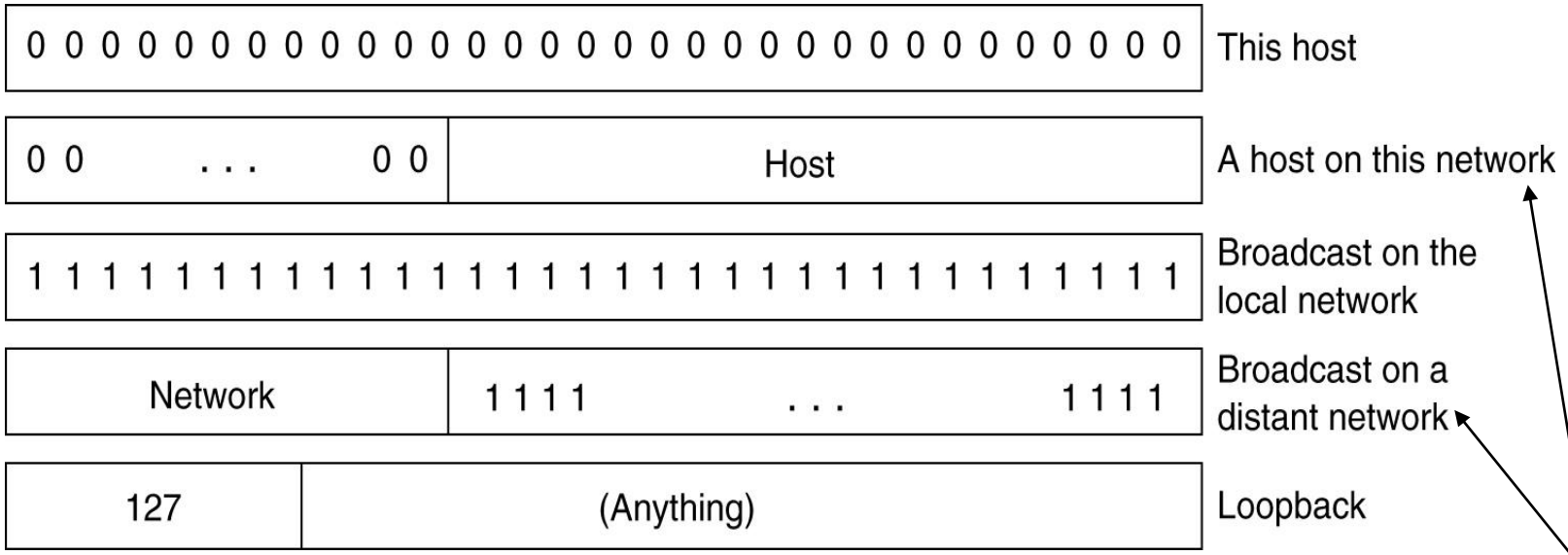
a) The network and host portions of the address are known by reading the <u>first bits</u>

b) All interfaces connected to the <u>same</u> **L3** network must have the **<u>same</u>** *network* portion

- Remember that a **<u>single L3</u>** network may consist of **<u>multiple L2</u>** networks connected via bridges/switches for example

IPv4 Addresses

Special IP addresses.



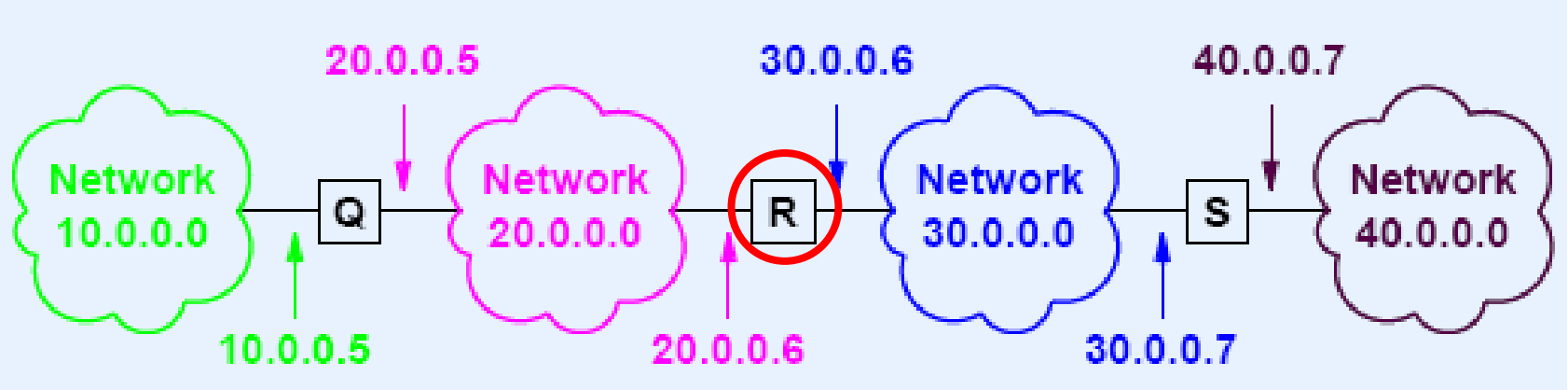
a) Other special IP addresses can be found in

http://www.iana.org/abuse/faq.html

IPv4 Address Conventions

- Conventionally, IPv4 addresses are written in dotted decimal notation
  - 0xC0290614 is written as 192.41.6.20
- The address of the host is written as the full 4 bytes
  - 192.168.2.2
- The address of the network is written with the host portions as zeros.
  - 128.46.0.0 refers to the network starting with 128.46
- The address length of the network portion of the address is added even if the class is known. Examples
  - 192.168.2.0/24
  - 128.46.0.0/16
- The network portion is called the ***prefix***
- Network address followed by all 1's in the host portion refers to **directed broadcast** to all hosts on the LAN
  - 192.168.2.255 refers to all hosts connected to the network 192.168.2.0
  - 128.46.255.255 refers to all hosts connected to the network 128.46.0.0

Basic Forwarding table

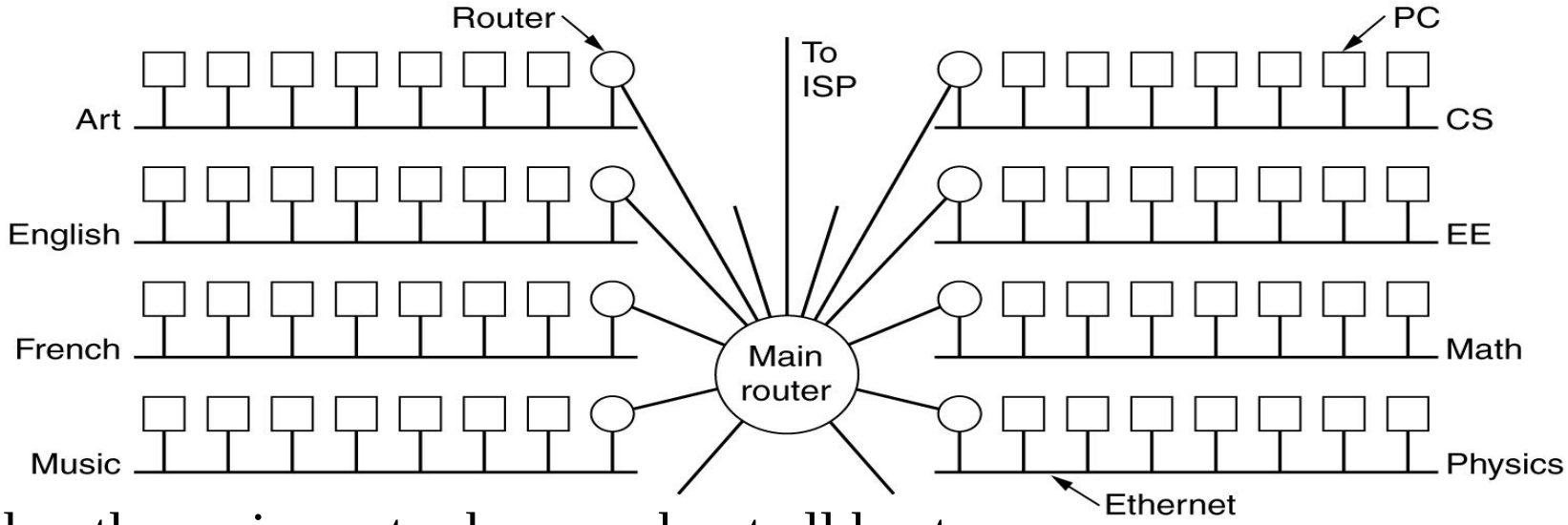


Forwarding table on router **R** looks as follows…

TO REACH NETWORK	ROUTE TO THIS ADDRESS
20.0.0.0 / 8	DELIVER DIRECT
30.0.0.0 / 8	DELIVER DIRECT
10.0.0.0 / 8	20.0.0.5
40.0.0.0 / 8	30.0.0.7

Subnets

- A campus network consisting of LANs for various departments.
- It can contain *thousands of hosts*

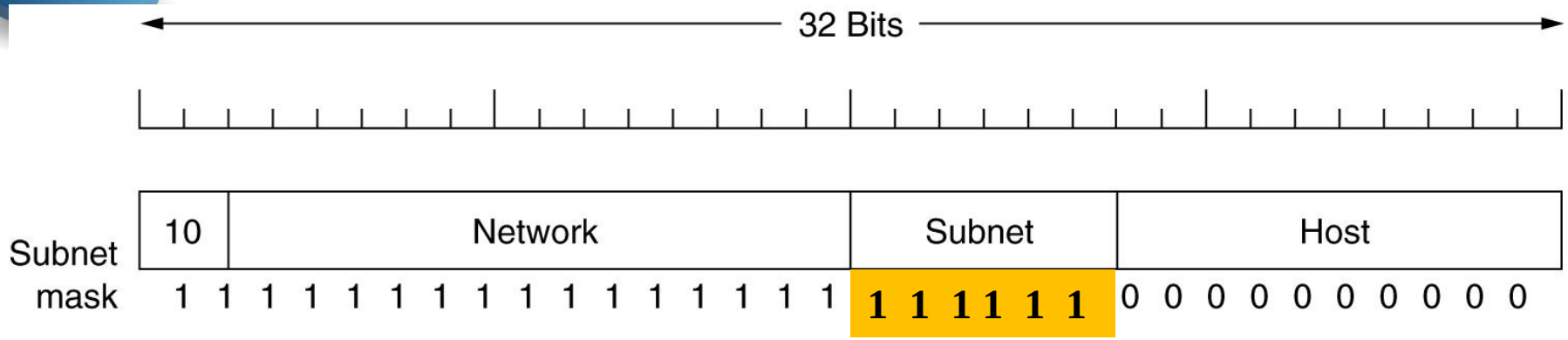


- Either the main router knows about <u>all hosts</u>

Or

- Allow the network to <u>split into several parts</u> but <u>appear as one part</u> to the **outside world**
- Subnetting splits up IP prefix to help with management

Subnets



- A class B network subnetted into 64 subnets ,
  We took **part of the host number** and used it to identify a **subnet**
- External routers need **not** know that we have divided our networks into subnets
- Each internal router in the network must have the subnet mask
  - Used to know the network, the subnet, and the host portions
- Each internal router needs to know about all the subnets in the network
- The subnet mask is written as
  - Dotted decimal 255.255.**252**.0
  - **/ notation**. In our example, /22

Subnets (Example)

- Consider the sub-netting of the class B address 130.50.0.0
- We want to divide it into three /22 subnets (we can increase it later)
  - Subnet 1: **10000010 00110010 000001**00 00000000
  - Subnet 2: **10000010 00110010 000010**00 00000000
  - Subnet 3: **10000010 00110010 000011**00 00000000
- The dotted decimal notation for the network address is
  - 130.50.4.0/22
  - 130.50.8.0/22
  - 130.50.12.0/22

Subnets Example: Class C

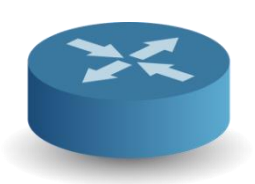
Let's use IP address **192.168.10.44** with subnet mask **255.255.255.248**or ***/29***. (29-24 = 5 bits, max 32 subnets with 6 hosts each with a total of 192 hosts)

IP Address (Decimal)	192.	168.	10.	44
IP Address (Binary)	11000000	10101000	00001010	00101100
Subnet Mask (Binary)	11111111	11111111	11111111	11111000
Subnet Mask (Decimal)	255.	255.	255.	248

IP Address (Decimal)	192.	168.	10.	44
IP Address (Binary)	11000000	10101000	00001010	00101100
Subnet Mask (Binary)	11111111	11111111	11111111	11111000
Subnet Address (Binary)	11000000	10101000	00001010	00101000
Subnet Address (Decimal)	192.	168.	10.	40

Bit wise AND

Subnet	0	8	16	…	40	…	248
First Host	1	9	17	…	41	…	249
Last Host	6	14	22	…	46	…	254
Broadcast	7	15	23	…	47	…	255

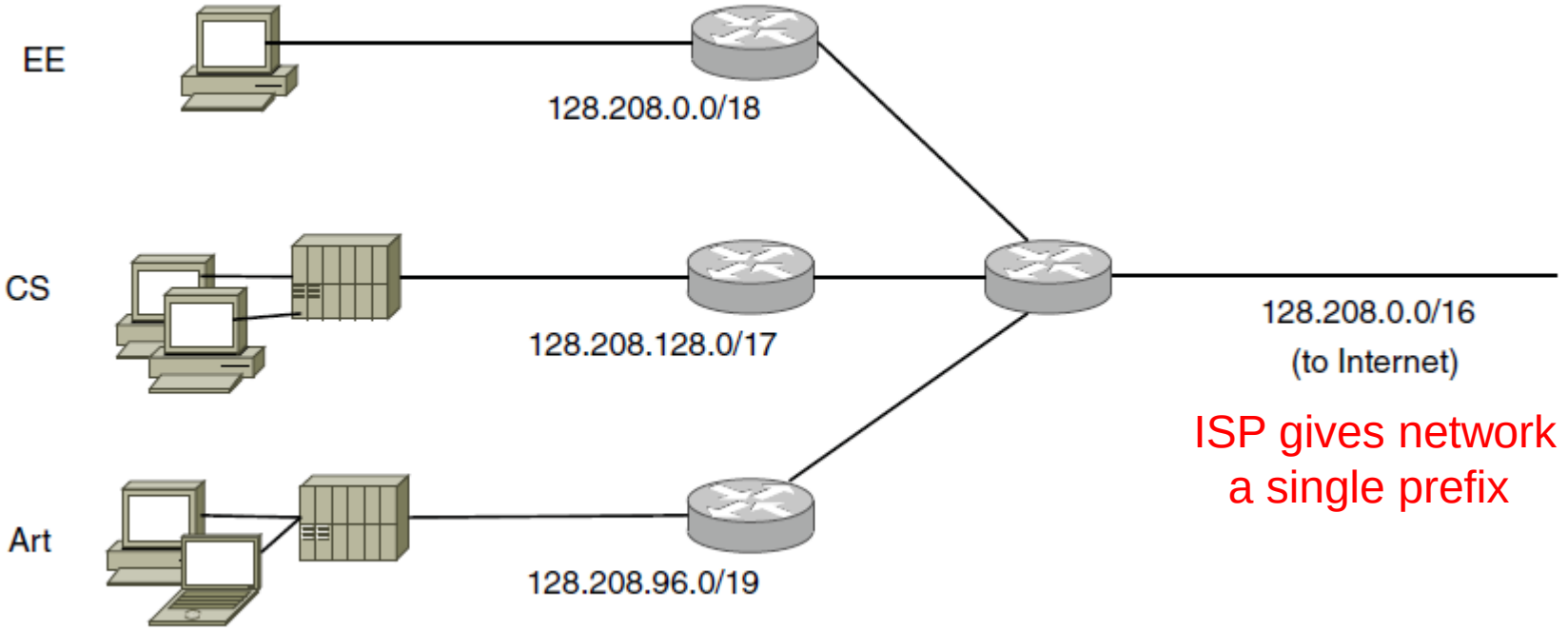


IP Addresses – Subnets

Subnetting splits up IP prefix to help with management

- Looks like a single prefix outside the network

Computer Science: 10000000 11010000 1|xxxxxxx xxxxxxxx
Electrical Eng.: 10000000 11010000 00|xxxxxx xxxxxxxx
Art: 10000000 11010000 011|xxxxx xxxxxxxx



ISP gives network a single prefix

Network divides it into subnets internally

How many addresses are left ?

CIDR – Classless InterDomain Routing rfc4632

Problems

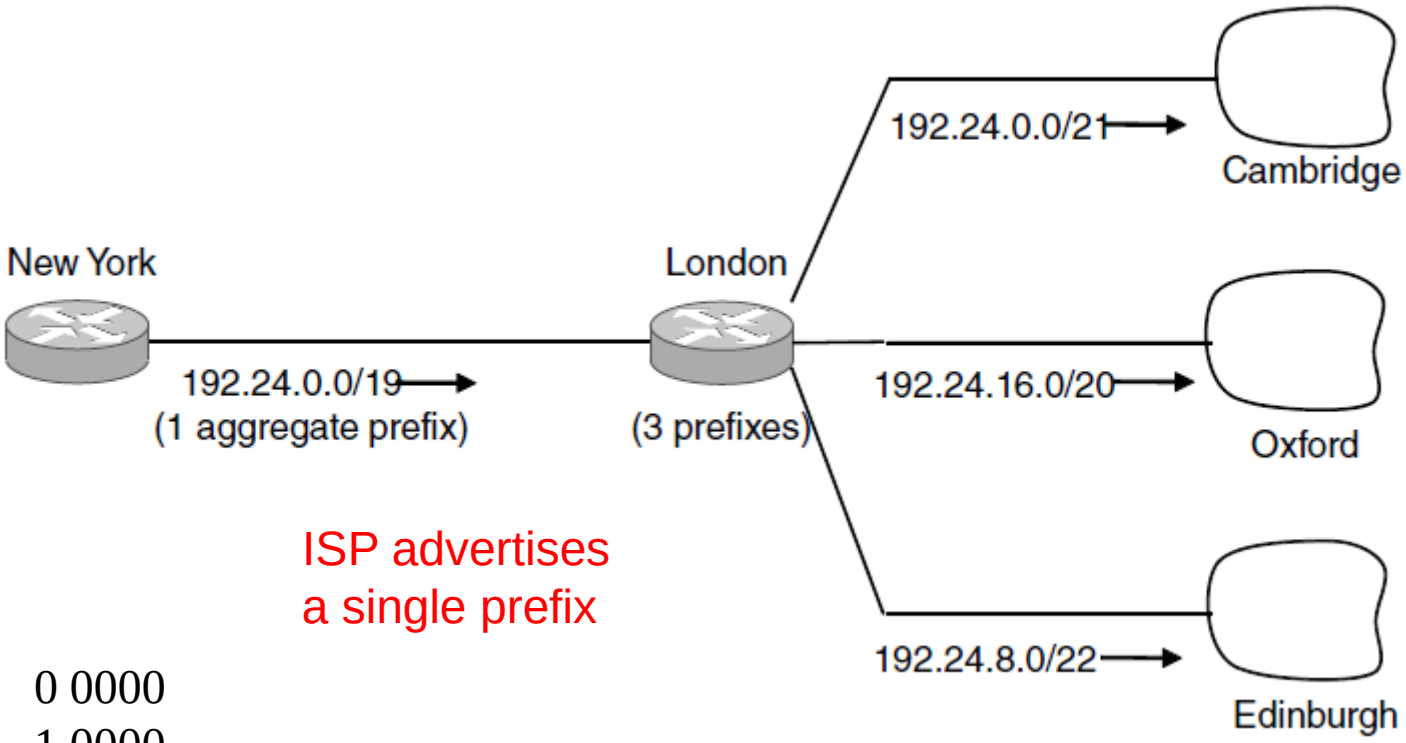
- Classes A and B addresses is too large for an organization (256M, 64k hosts)
- There are only 254 class A and 16K class B addresses
- Class C is too small for an organization (only 254)
- Assigning multiple class C addresses to the same organization increases routing table in the core

Solution

- Forget the classful addressing
- All routers must carry a prefix and a mask
- IP address of an interface MUST have a subnet mask
  - Because there is no classful addressing, the only way to know the network and host portions of an address is to configure a subnet mask with the IP address
- Send pair of **(address, mask)** whenever exchanging topology information
- SometimeS the address/mask pair is called a **CIDR block**

IP Addresses – Aggregation

Aggregation joins multiple IP prefixes into a single larger prefix to reduce routing table size



ISP advertises a single prefix

ISP customers have different prefixes

17 18 19

0 0 0 0 0000

0 0 0 1 0000

0 0 0 0 1000

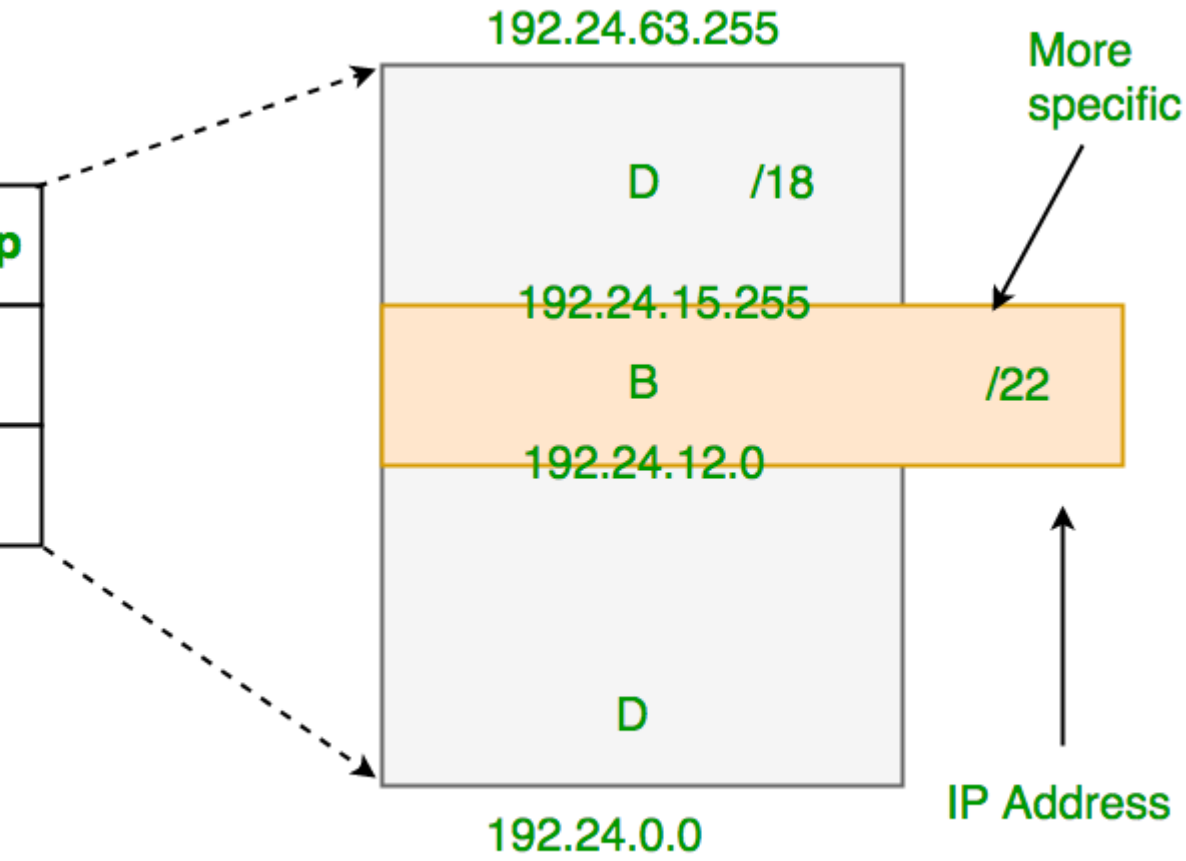
Forwarding for CIDR

a) Each entry in the forwarding table contains prefix/mask

b) Perform a **longest-prefix** match
- Bitwise-AND the incoming IP address with mask in the entry
- If the resulting bitwise-AND operation matches multiple entries, choose the entry that has the longest prefix length
- If the resulting bitwise-AND does not match any entry
  - If there is a default entry, use it
  - Otherwise, drop the packet

c) A prefix P1 with prefix length L1 is said to be **less specific** than P2 with prefix length L2, if
- L1 < L2
- The first L1 bits in P1 and P2 are identical
- Example 128.10.0.0 /16 is less specific than 128.10.2.0/24
- Sometimes we say : P1 **covers** P2
- Sometimes we say P1 is an **aggregate** route of P2

Prefix	Next Hop
192.24.0.0/18	D
192.24.12.0/22	B



Exercise

Classless Inter-domain Routing (CIDR) receives a packet with address **131.23.151.76.**

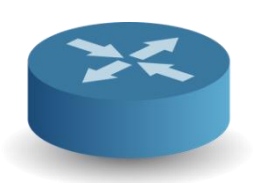
The router's routing table has the following entries:

Prefix	Output Interface Identifier
131.16.0.0/12	3
131.28.0.0/14	5
131.19.0.0/16	2
131.22.0.0/15	1

The identifier of the output interface on which this packet will be forwarded is _____.

Answer: "1". **23: 0001 0111 / 16: 0001 0000 / 22: 0001 0110 / 28: 0001 1100 / 19: 0001 0011**

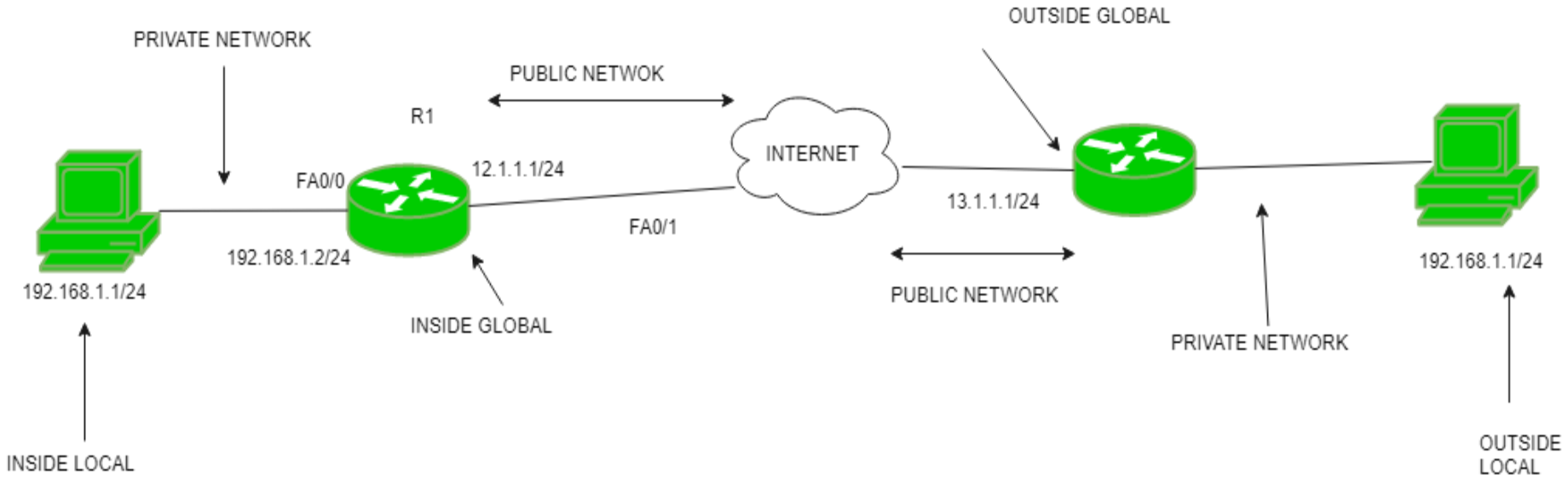
We need to first find out matching table entries for incoming packet with address "131.23.151.76". The address matches with two entries "131.16.0.0/12" and "131.22.0.0/15" (We found this by matching first 12 and 15 bits respectively).
So should the packet go to interface 3 or 1? We use Longest Prefix Matching to decide among two. The most specific of the matching table entries is used as the interface. Since "131.22.0.0/15" is most specific, the packet goes to interface 1.



IP Addresses – NAT

- To access the Internet, one public IP address is needed, but we can use a private IP address in our private network. The idea of NAT is to allow multiple devices to access the Internet through a single public address. To achieve this, the translation of private IP address to a public IP address is required
- **Network Address Translation (NAT)** is a process in which one or more local IP address is translated into one or more Global IP address and vice versa in order to provide Internet access to the local hosts.
- It does the translation of port numbers i.e. masks the port number of the host with another port number, in the packet that will be routed to the destination. It then makes the corresponding entries of IP address and port number in the NAT table.

NAT illustration

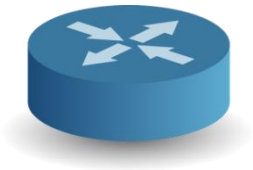


Inside local address – An IP address that is assigned to a host on the Inside (local) network. The address is probably not a IP address assigned by the service provider i.e., these are private IP address. This is the inside host seen from the inside network.

Inside global address – IP address that represents one or more inside local IP addresses to the outside world. This is the inside host as seen from the outside network.

Outside local address – This is the actual IP address of the destination host in the local network after translation.

Outside global address – This is the outside host as seen form the outside network. It is the IP address of the outside destination host before translation.



NAT

The three reserved ranges are:

10.0.0.0 – 10.255.255.255/8 (16,777,216 hosts)

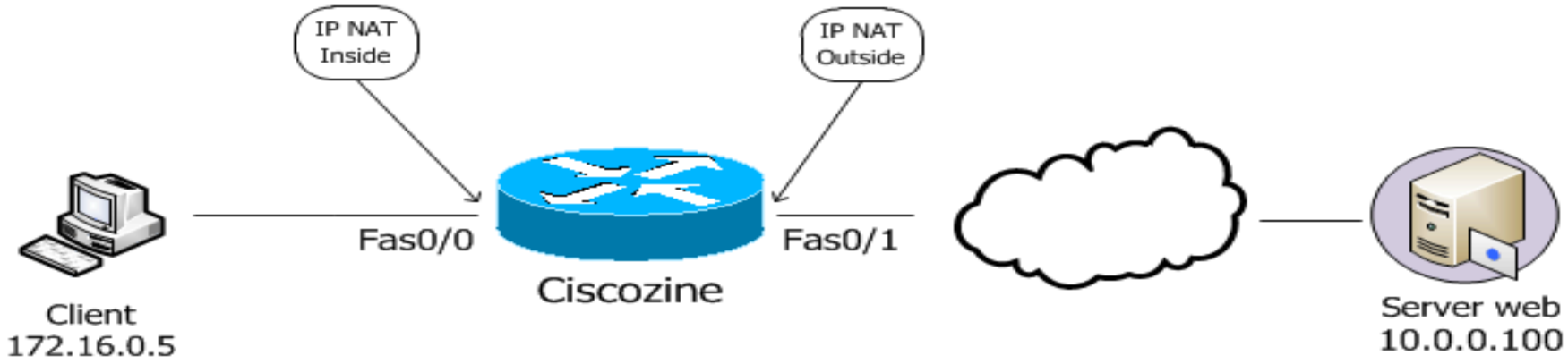
172.16.0.0 – 172.31.255.255/12 (1,048,576 hosts)

192.168.0.0 – 192.168.255.255/16 (65,536 hosts)

Read in Tanenbaum page 454-455 for disadvantages of NAT

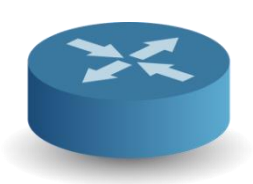
Simple NAT

Static source NAT



The source IP address is translated when the packets leave the outside interface (Fa0/1)

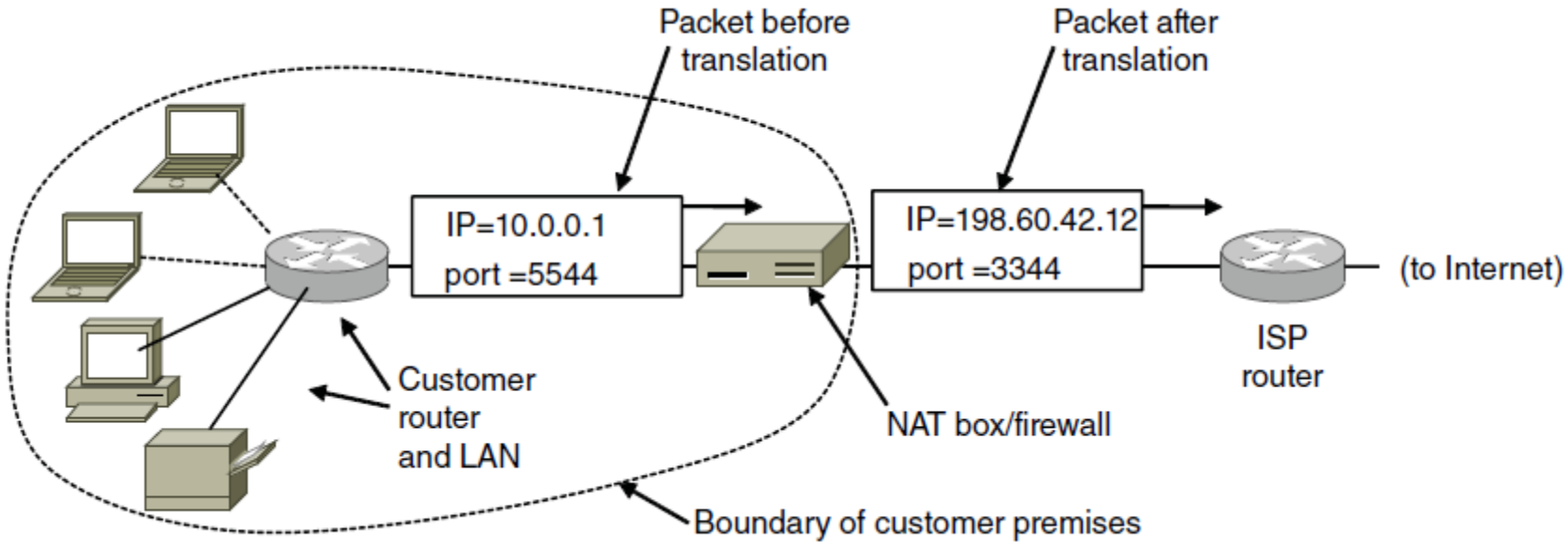
IP SRC: 172.16.0.5 IP DST: 10.0.0.100	IP SRC: 10.16.0.5 IP DST: 10.0.0.100



Port Address Translation PAT

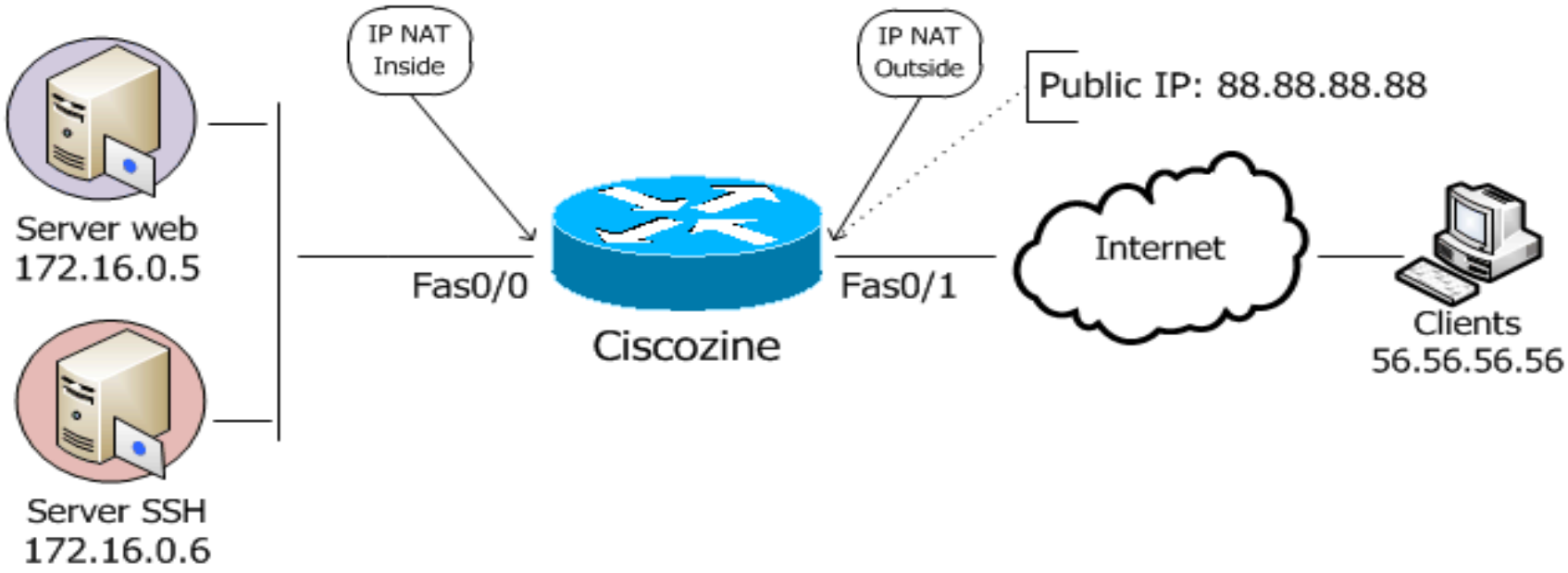
NAT (Network Address Translation) box maps one external IP address to many internal IP addresses

- Uses TCP/UDP port to tell connections apart
- Violates layering; very common in homes, etc.



PAT Example

Static PAT



IP SRC: 56.56.56.56
IP DST: 172.16.0.5 tcp 80

IP SRC: 56.56.56.56
IP DST: 88.88.88.88 tcp 80

IP SRC: 56.56.56.56
IP DST: 172.16.0.6 tcp 22

IP SRC: 56.56.56.56
IP DST: 88.88.88.88 tcp 666

Internet Control Protocols

IP works with the help of several control protocols:

- ICMP is a companion to IP that returns error info
  - Required, and used in many ways, e.g., for traceroute
- ARP finds Ethernet address of a local IP address
  - Glue that is needed to send any IP packet
  - Host queries an address and the owner replies
- DHCP assigns a local IP address to a host
  - Gets host started by automatically configuring it
  - Host sends request to server, which grants a lease

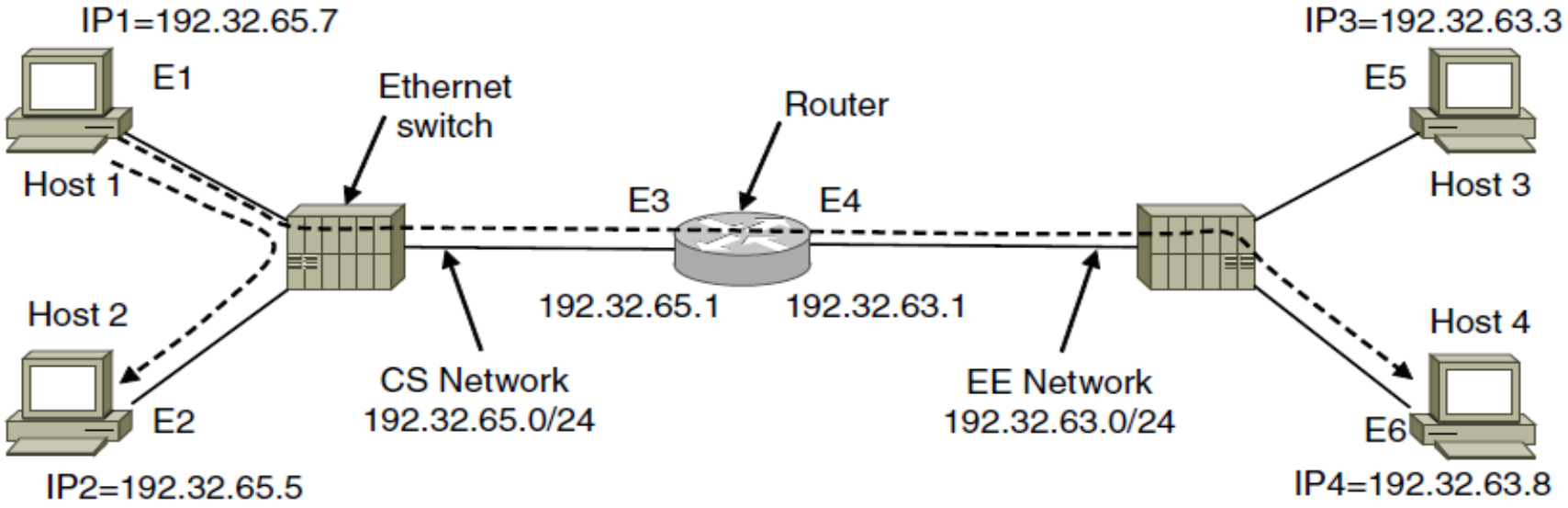
Internet Control Protocols: ICMP

Main ICMP (Internet Control Message Protocol) types:

Message type	Description
Destination unreachable	Packet could not be delivered
Time exceeded	Time to live field hit 0
Parameter problem	Invalid header field
Source quench	Choke packet
Redirect	Teach a router about geography
Echo and Echo reply	Check if a machine is alive
Timestamp request/reply	Same as Echo, but with timestamp
Router advertisement/solicitation	Find a nearby router

Internet Control Protocols: ARP

ARP (Address Resolution Protocol) lets nodes find target Ethernet addresses [pink] from their IP addresses

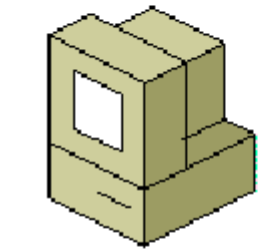


Frame	Source IP	Source Eth.	Destination IP	Destination Eth.
Host 1 to 2, on CS net	IP1	E1	IP2	E2
Host 1 to 4, on CS net	IP1	E1	IP4	E3
Host 1 to 4, on EE net	IP1	E4	IP4	E6

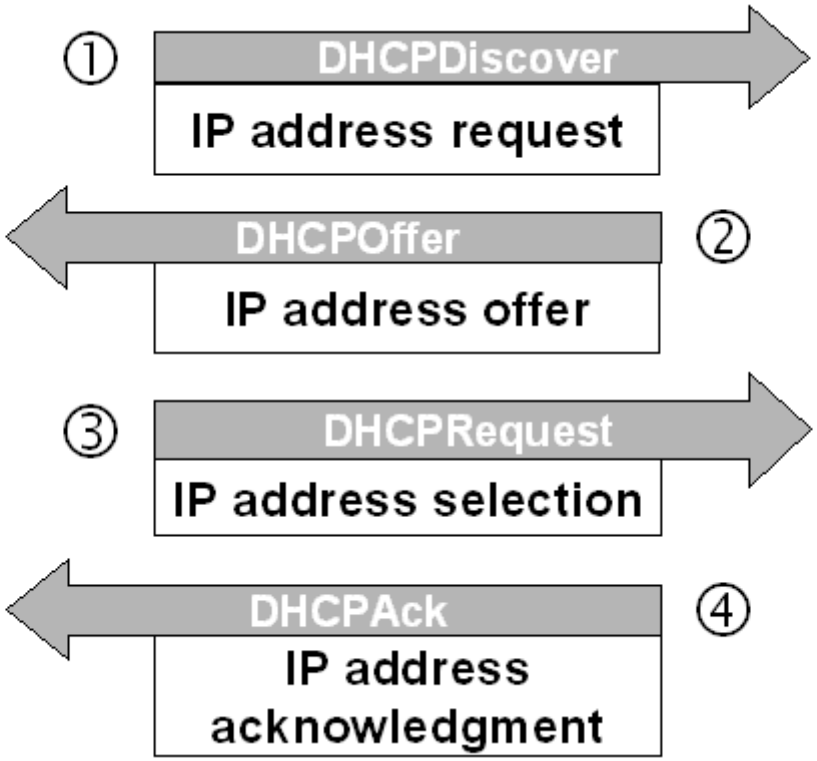
Internet Control Protocols: DHCP

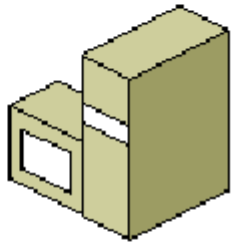
- The Dynamic Host Configuration Protocol (DHCP) can help with the workload of configuring systems on a network by assigning addresses to systems on boot-up automatically.
- It also provides a central database of devices that are connected to the network and eliminates duplicate resource assignments
- DHCP provides an automated way to distribute and update IP addresses and other configuration information on a network. A DHCP server provides this information to a DHCP client through the exchange of a series of messages, known as the DHCP conversation or the DHCP transaction

DHCP Messages



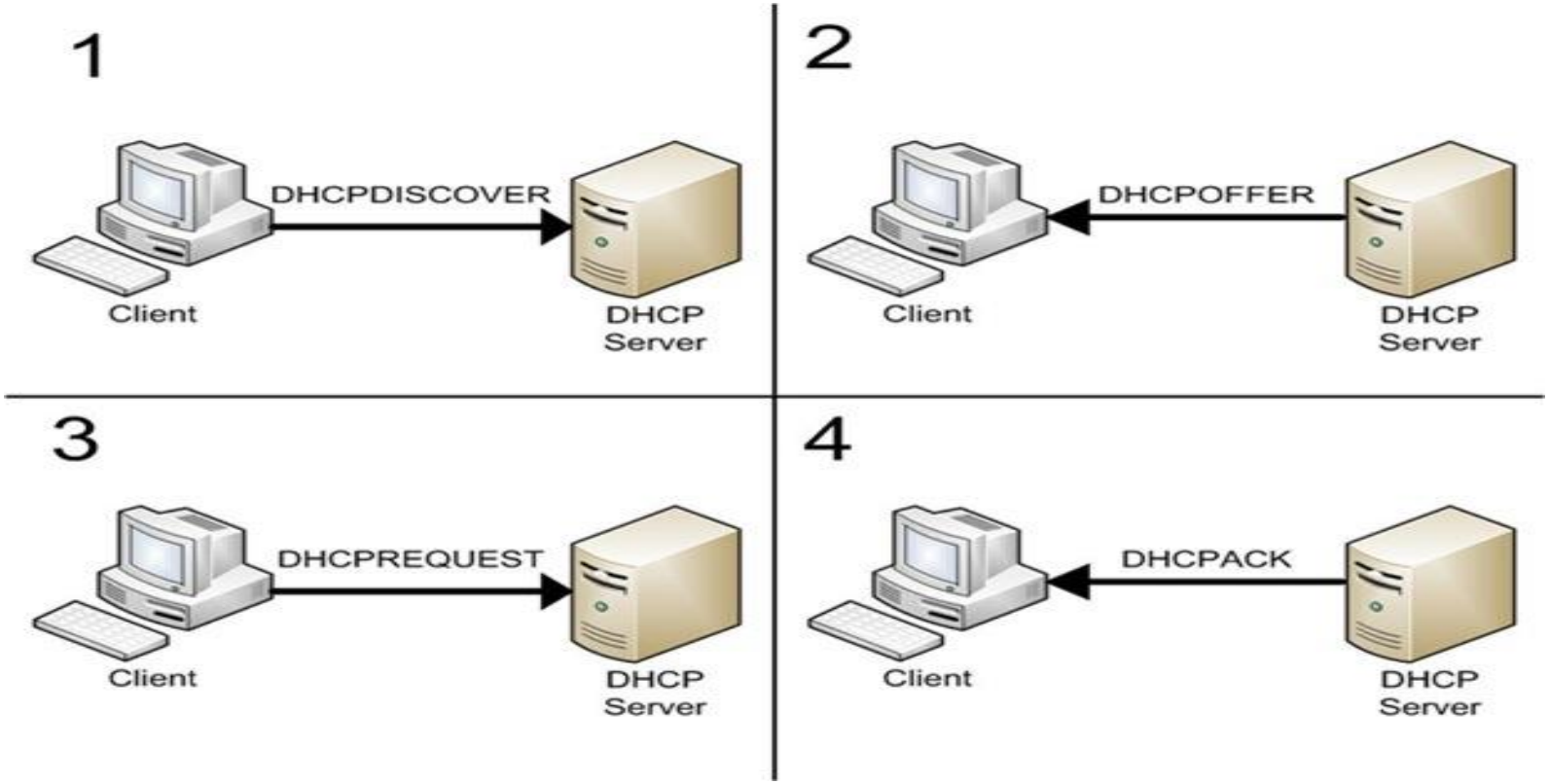
DHCP client





DHCP server

DHCP Scenario



The acknowledgement phase involves sending a DHCPACK packet to the client. This packet includes the lease duration and any other configuration information that the client might have requested. At this point, the IP configuration process is completed

IP Version 6 (1)

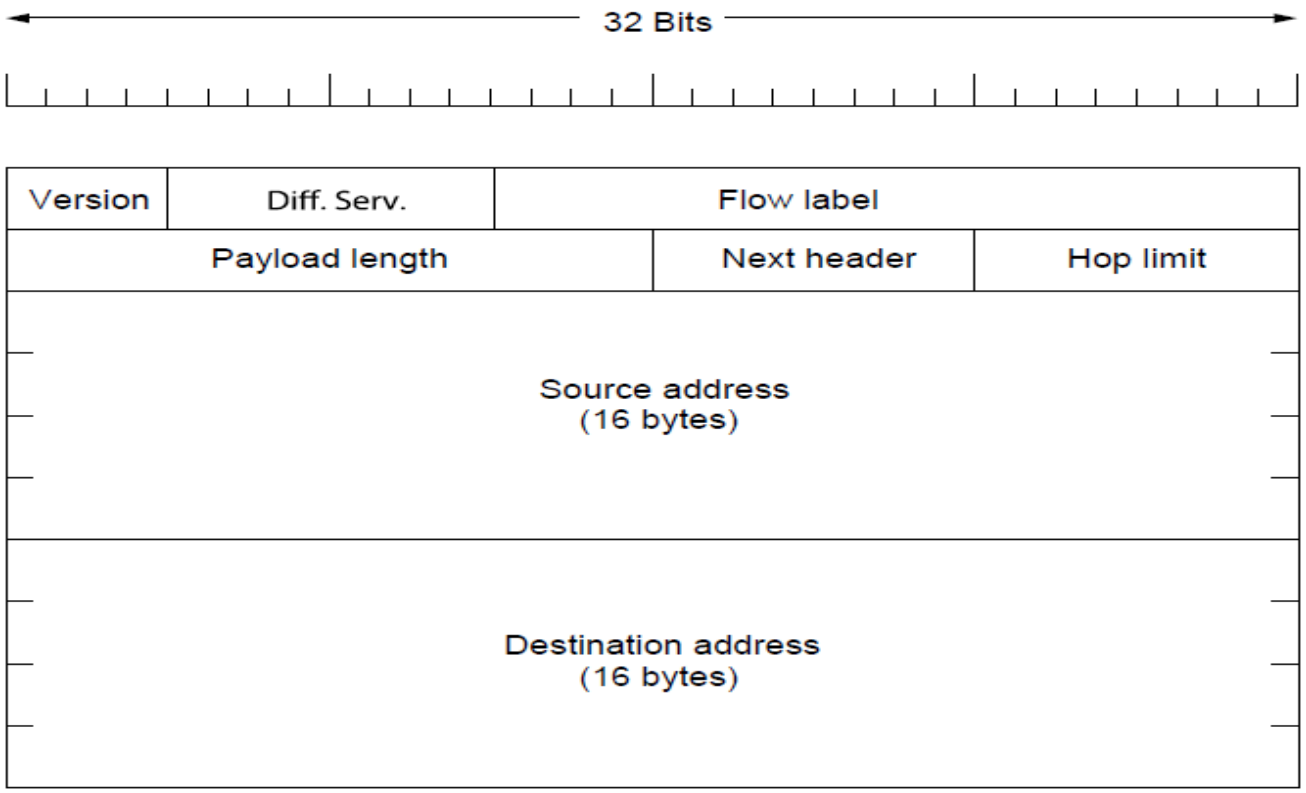
Major upgrade in the 1990s due to impending address exhaustion, with various other goals:

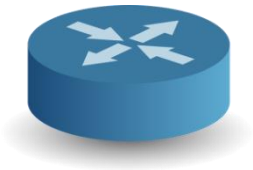
- Support billions of hosts
- Reduce routing table size
- Simplify protocol
- Better security
- Attention to type of service
- Aid multicasting
- Roaming host without changing address
- Allow future protocol evolution
- Permit coexistence of old, new protocols, …

Deployment has been slow & painful, but may pick up pace now that addresses are all but exhausted

IP Version 6 (2)

IPv6 protocol header has much longer addresses (128 vs. 32 bits) and is simpler (by using extension headers)





IP Version 6 (3)

IPv6 6 extension headers handles other functionality

Extension header	Description
Hop-by-hop options	Miscellaneous information for routers
Destination options	Additional information for the destination
Routing	Loose list of routers to visit
Fragmentation	Management of datagram fragments
Authentication	Verification of the sender's identity
Encrypted security payload	Information about the encrypted contents