NANYANG TECHNOLOGICAL UNIVERSITY

CE3005: Computer Networks

Module 2-2:
Network Layer - Internet Protocol (IP)

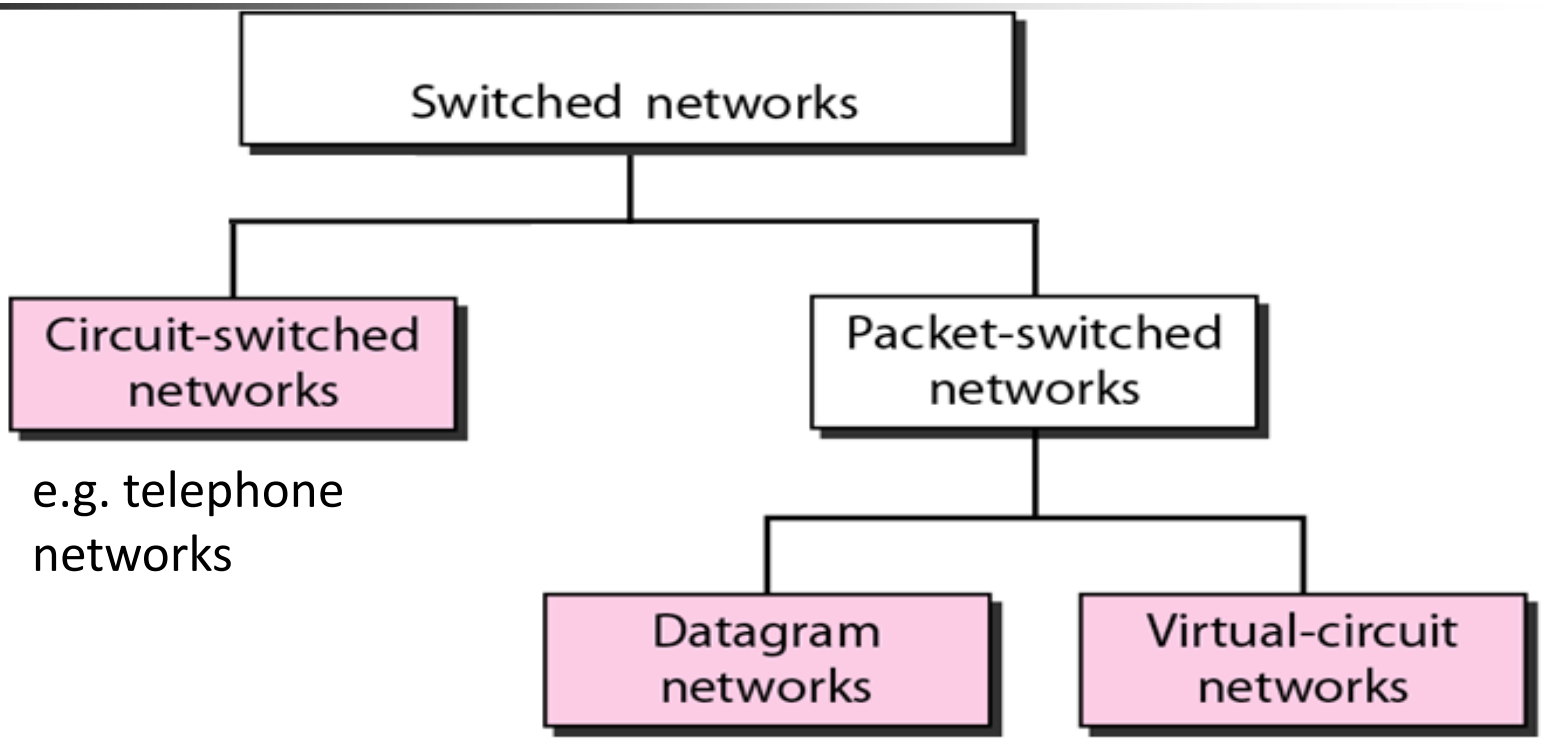
Semester 1 2016-2017

School of Computer Engineering

Contents

- IP Functions: Addressing and Fragmentation
- IPv4 Addressing
  - Classful Addressing
  - Subnet and Subnet Mask
- IPv4 Address Exhaustion
  - CIDR Addressing
  - Network Address Translation (NAT)
  - IPv6: version 6
- IP Fragmentation and Reassembly
- IP Routing and ARP
- Internet Control Message Protocol (ICMP)
  - ping utility
  - tracert utility

In theory, network layer may be implemented using circuit or packet-switching, which can further be divided into datagram or virtual-circuit networks.



e.g. telephone networks

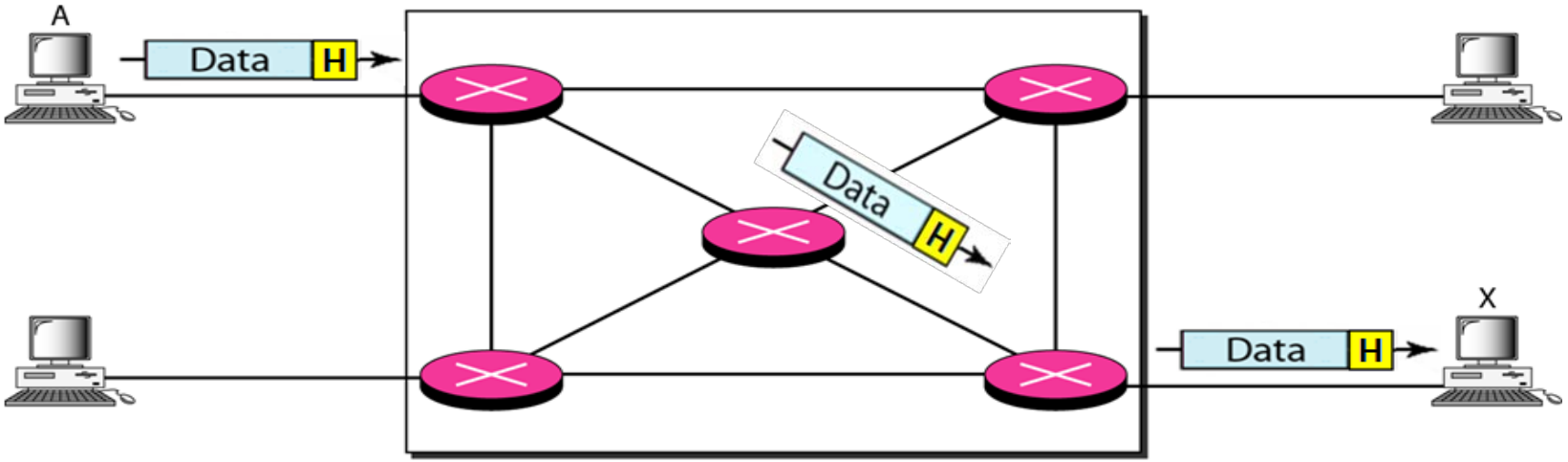
e.g. IP networks

e.g. X.25, ATM, frame relay

We will focus on packet-switched, specifically datagram network, which is the basis for the Internet technology.

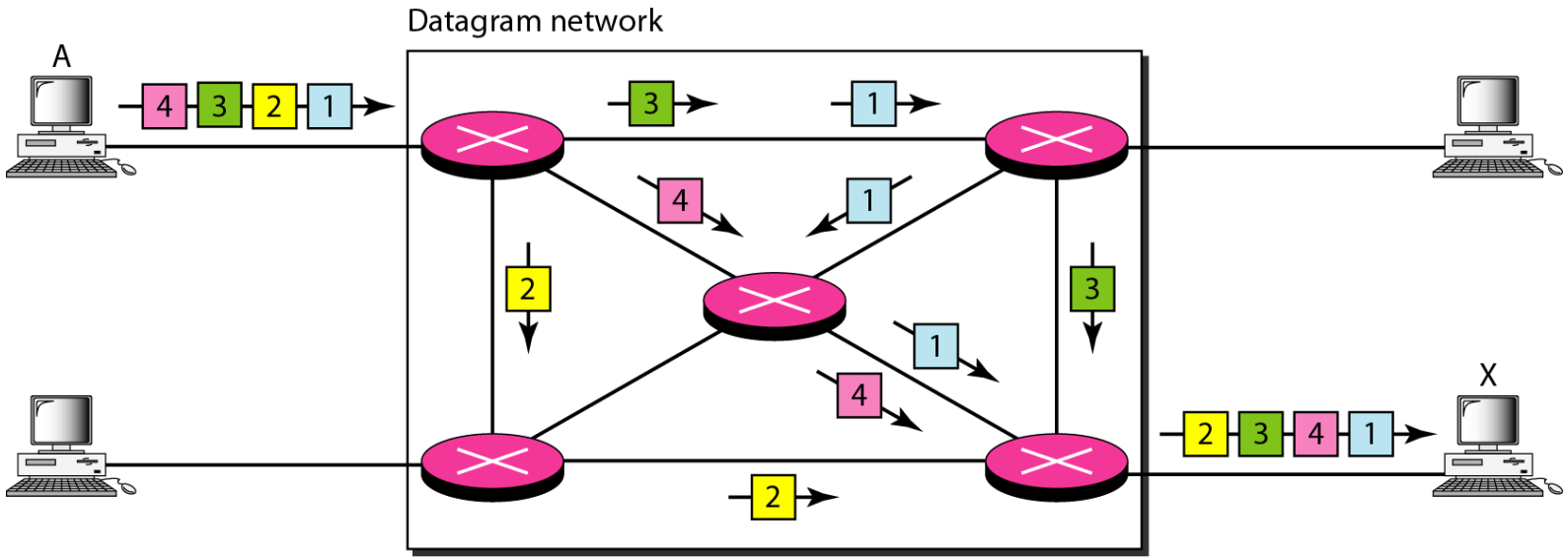
In packet-switched datagram networks, transmission begins without establishing a communication path – Addressing is needed in each packet.

For example, to transmit data from A to X:



1. A can transmit immediately;
2. However, A needs to append a header H containing the destination address so that intermediate packet-switched nodes know how to send it to X;
3. At the same time, packet-switched nodes are shared with others for transmissions (efficient use of resources).

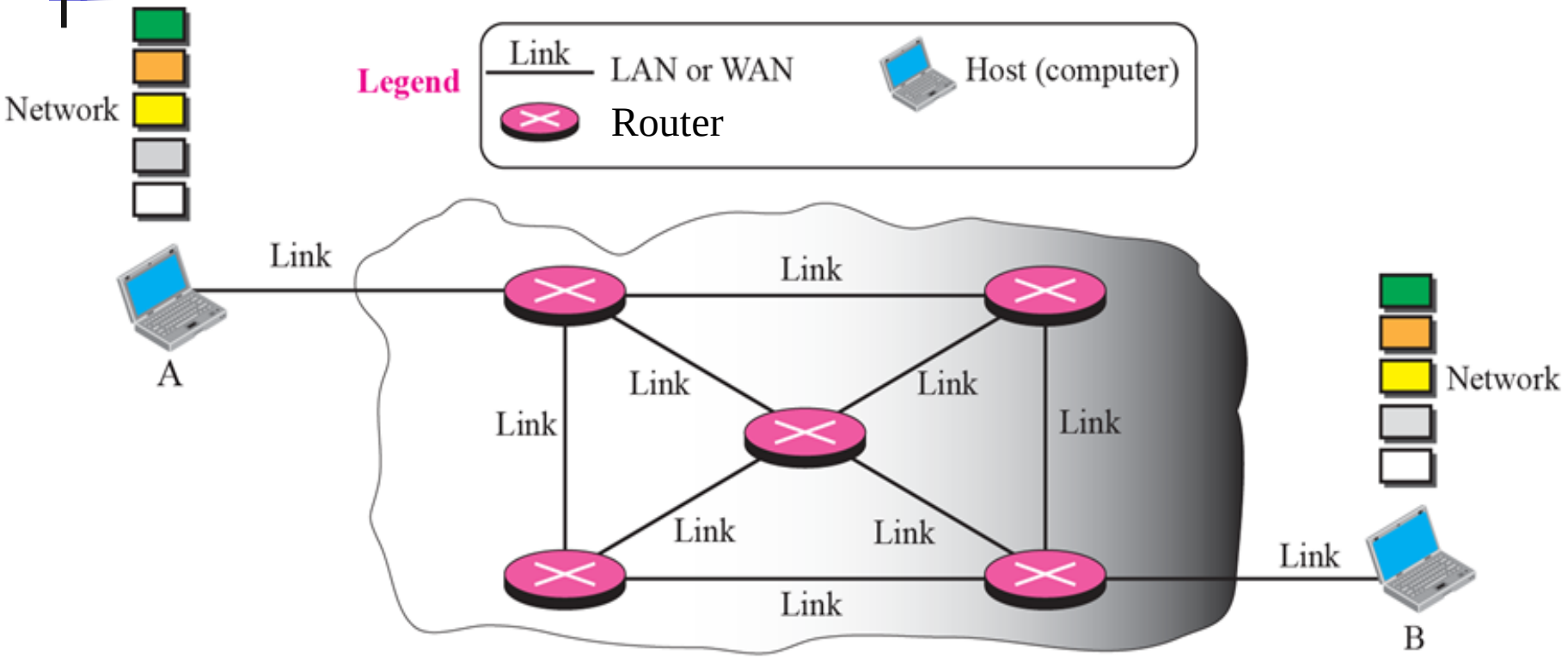
In addition, data is divided into suitable size packets depending on MTU (Maximum Transfer Unit) of individual network - Fragmentation.



Note that since there is no reservation of communication path, different packets may take different path depending on the load of packet switches.

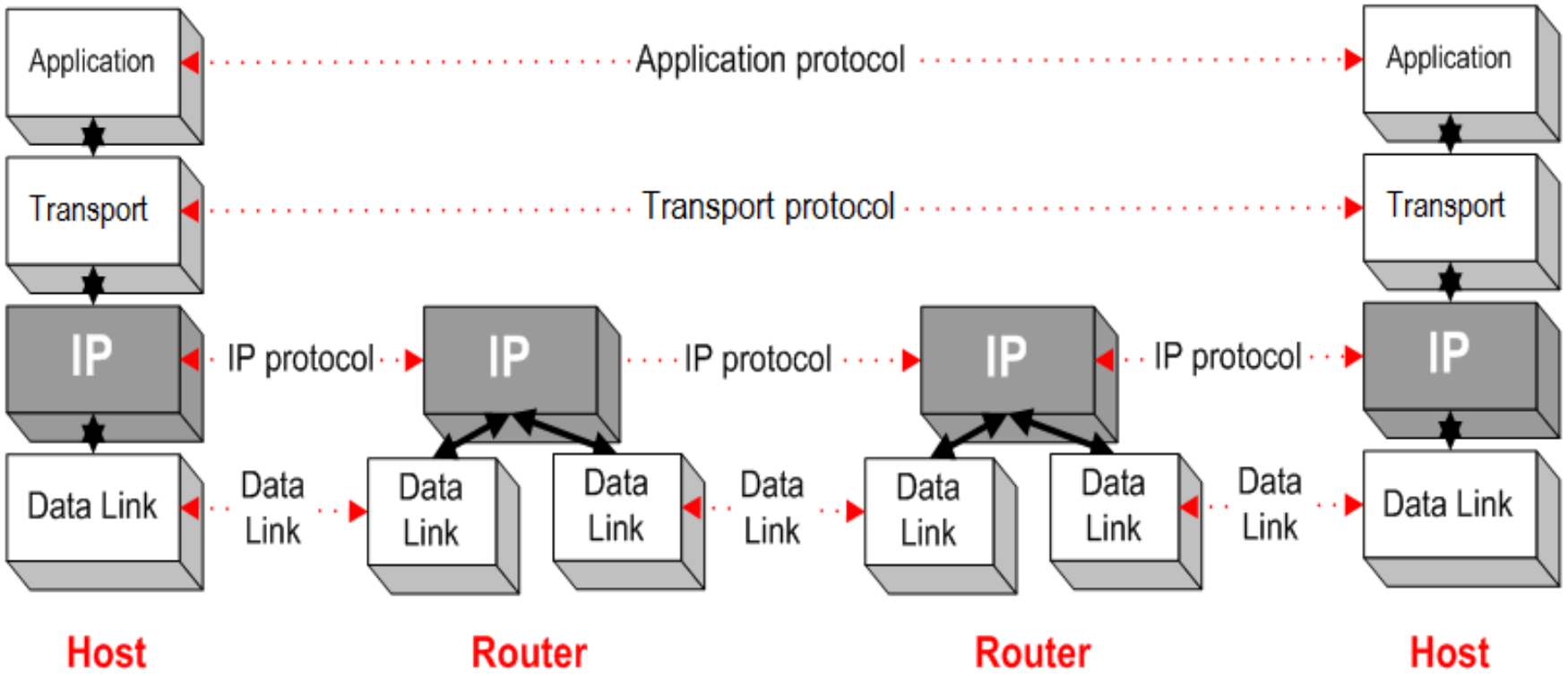
As a result, packets may arrive out of sequence, or even corrupted and lost! (It will be up to upper layer to handle this, e.g. TCP at transport layer – to be discussed in later module.)

To enhance our understanding of packet-switched network, let's study the operations of a network layer protocol - the Internet Protocol (IP).



Note: Internet may be viewed as consisting of links (LANs and WANs) interconnected together by packet switches nodes called routers.

Internet Protocol (IP) is implemented at both hosts and routers.



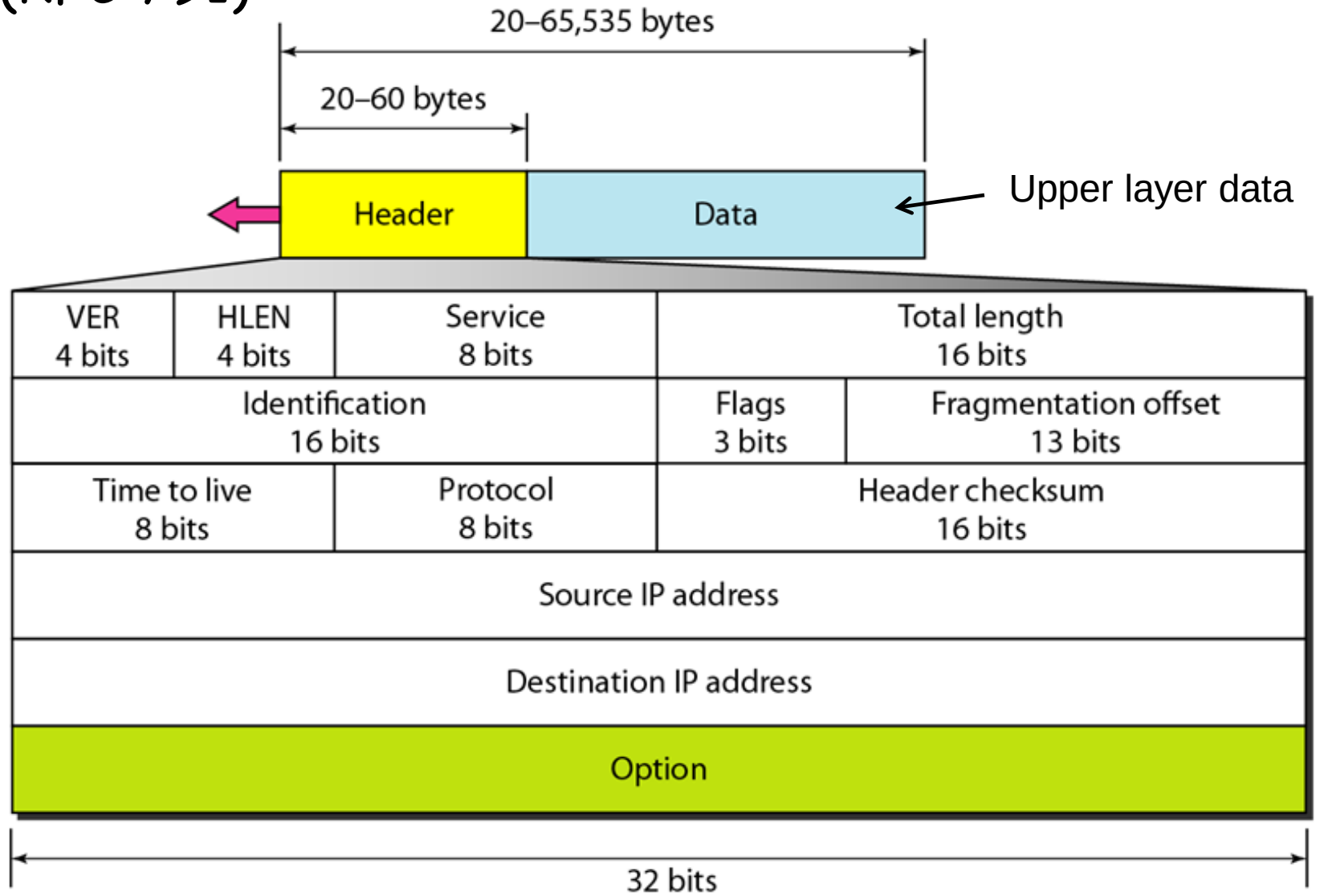
A packet will traverse through intermediate routers hop-by-hop from source to destination.

Internet Protocol

Some characteristics:

- **It has two basic protocol functions:**
  - **Addressing**
  - **Fragmentation**
- **It provides a connectionless unreliable best-effort (datagram) service:**
  - **Connectionless: each packet is handled independently, no flow control**
  - **Unreliable: no error control**
  - **Best-effort: no throughput guarantee, no delay guarantee, no Quality of Service (QoS) guarantee**

To support the required functions, the IP header, specifically IPv4 (version 4), is designed as follows: (RFC 791)



IPv4 Header

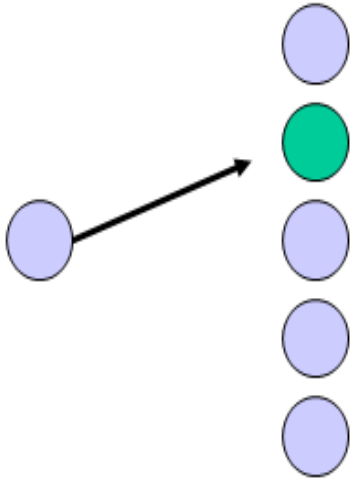
- **Version (VER):** Version number of IP. Current widely-used version is 4.
- **Internet Header Length (HLEN):** Length of header (in multiples of 4 bytes). Typically 5, representing header length of 5x4 = 20 bytes.
- **Type of Service:** ignore
- **Total Length:** Length of datagram including header (in bytes). Max. is 65,535.
  - Max typical size -1,500, Jumbo size - 9,216
- **Option Fields (variable length):** ignore

IPv4 Header

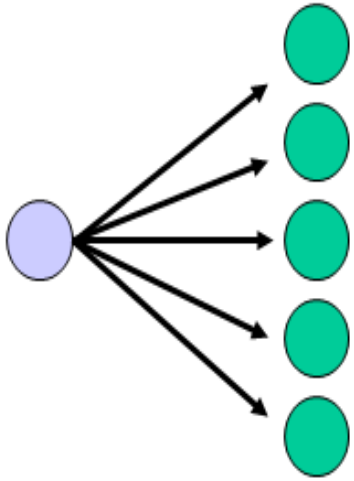
- **Protocol: Indicates the protocol that IP is carrying.** (01_{16} for ICMP, 06_{16} for TCP, 11_{16} for UDP)
- **Header Checksum: For verifying the header is free from error - ignore.**
- **Source and Destination IP Address: Indicates the IP addresses of source and destination.**

NOTE: Remaining fields will be discussed in relevant slides later.

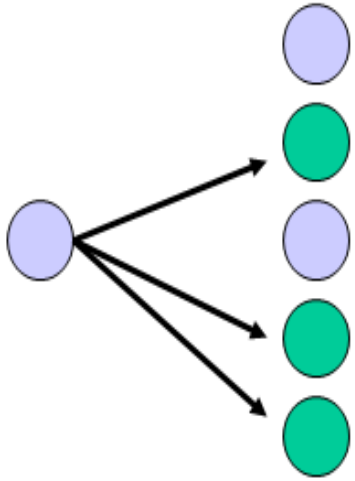
IP Addressing



Unicast:
one-to-one

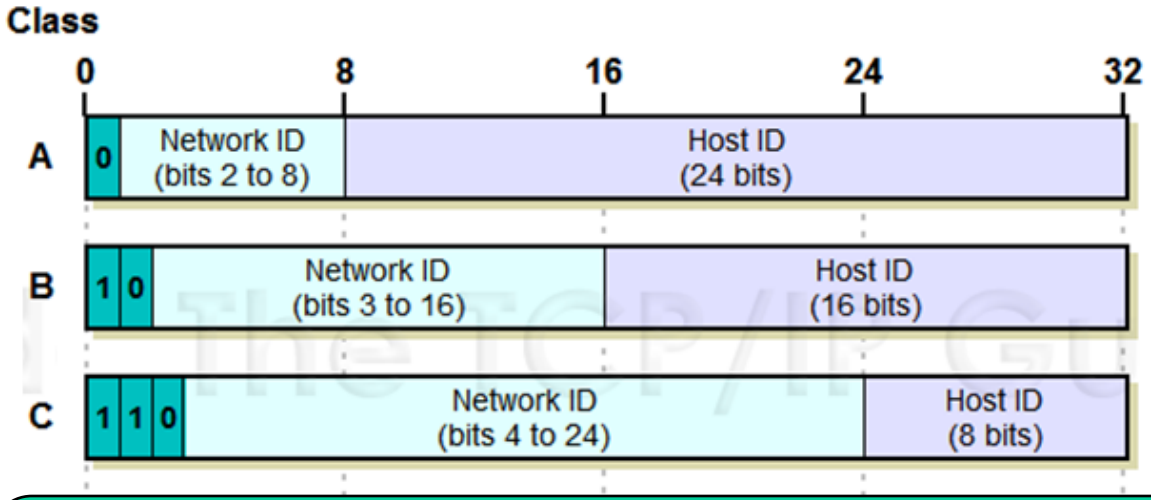


Broadcast:
one-to-all



Multicast:
one-to-many

IP addresses, specifically IPv4, are 32-bit and are traditionally divided into 5 classes; hence also called classful addressing.

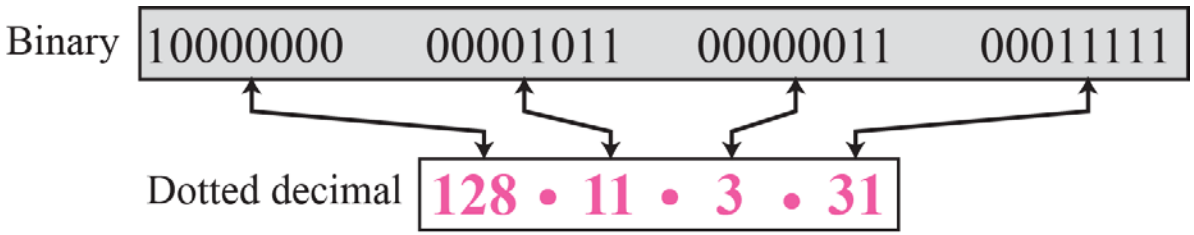


1.0.0.0 to 126.255.255.255

128.0.0.0 to 191.255.255.255

192.0.0.0 to 223.255.255.255

IP address notation:

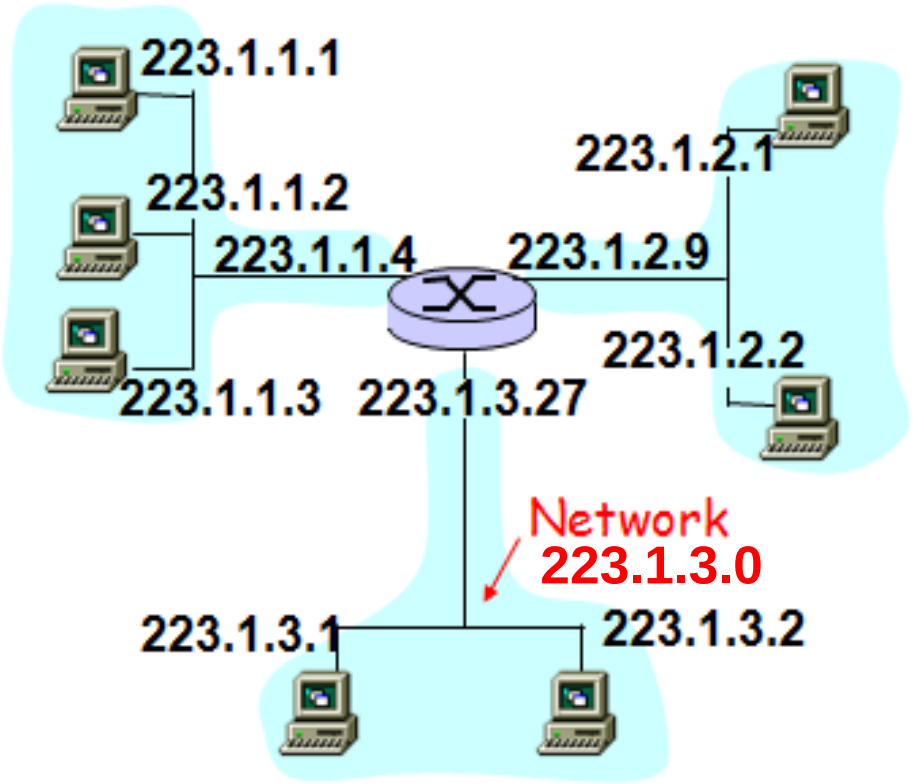


IPv4 Address: Special Use (RFC 5735)

- Network and/or Host id – all '0's: (can only be used as source address; e.g. during startup to get own IP in DHCP)
  - 0.0.0.0 → means this host on this network.
  - 0.0.0.10 → host 10 on this network.
- Network and/or Host id – all '1's: (can only be used as destination address)
  - 255.255.255.255 → limited broadcast within this network. (ARP)
  - 155.69.255.255 → directed broadcast on 155.69.x.x network.
- Loopback Address (127.x.y.z):
  - Internal loopback to same host. Useful for self-testing of network software. "x.y.z" can be any valid value, eg, 127.0.0.1.

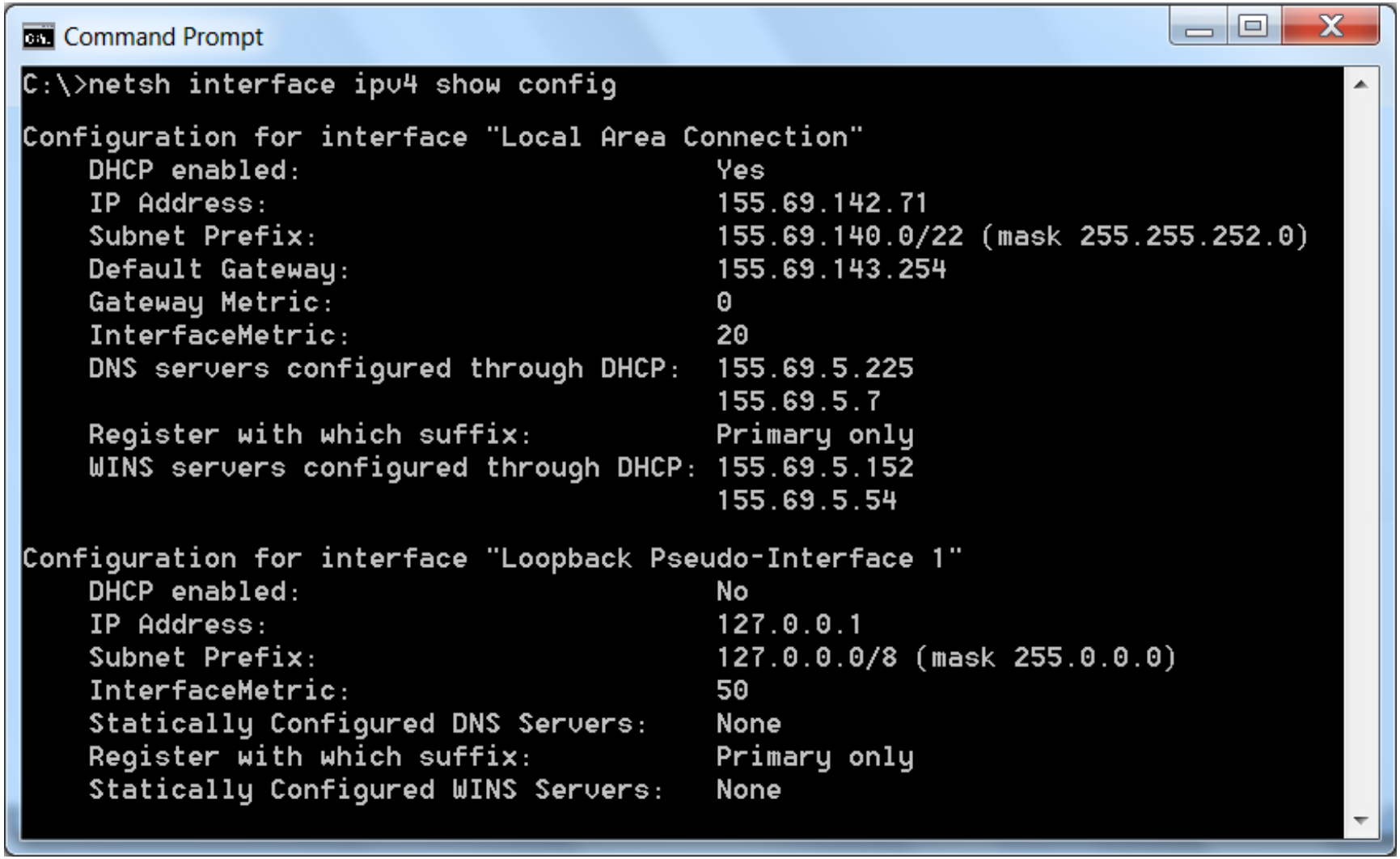
IPv4 Address

- **IP address: assigned to host/router's *interface***
- **Interface: connection between host/router and physical link**
- **What's a network? (from IP address perspective)**
  - **device interfaces with same network id of IP address**
  - **can physically reach each other without intermediate router**



Note: a host/router may have multiple interfaces, and hence multiple IP addresses.

Example of interfaces in a Windows host:



Note: netsh command seems to work only in Windows Vista/7.

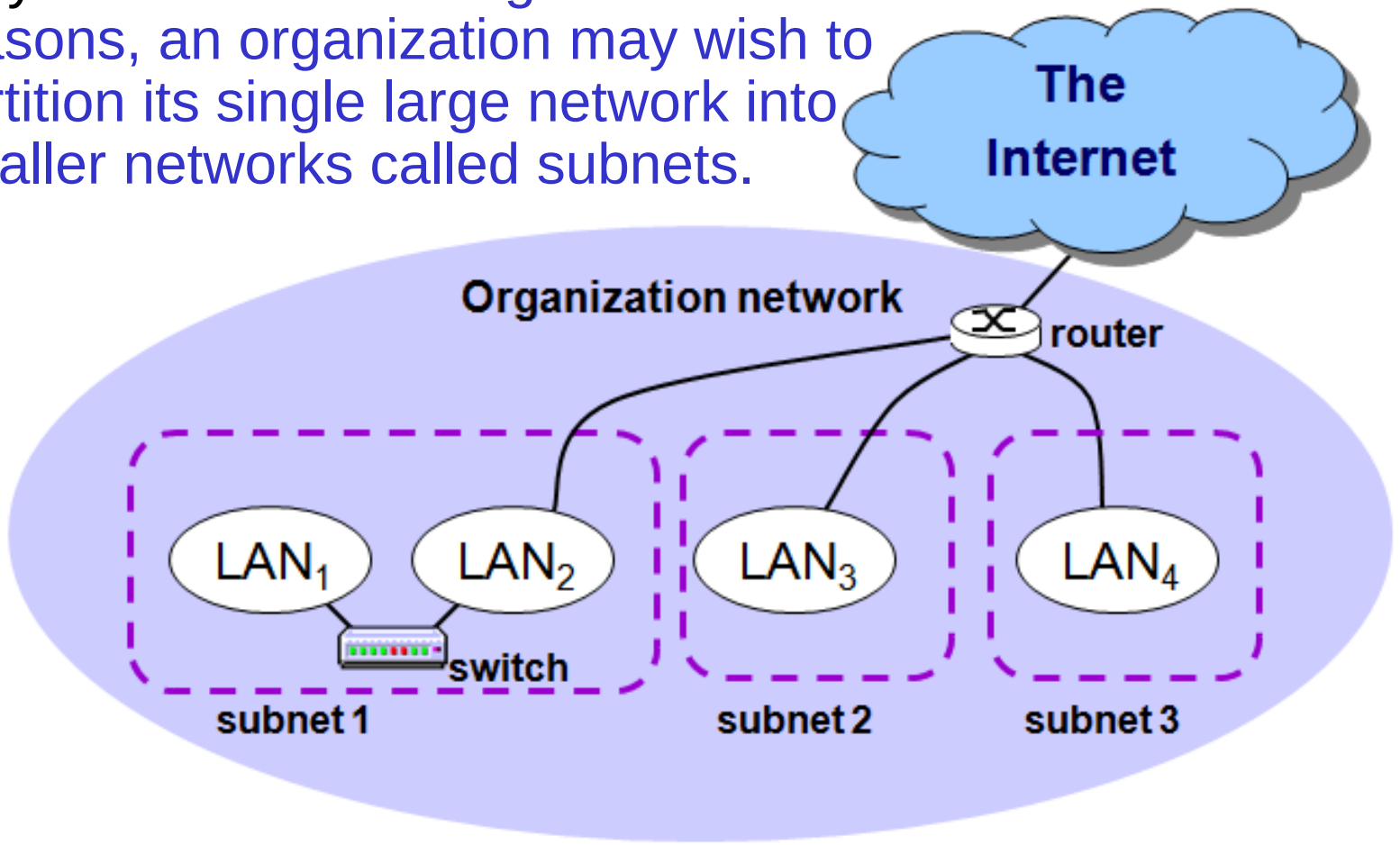
Ethernet adapter Local Area Connection:

```
   Connection-specific DNS Suffix  . : ntu.edu.sg
   Description . . . . . . . . . . . : Intel(R) 82567LM-3 Gigabit Network Connec
tion
   Physical Address. . . . . . . . . : 00-25-11-89-50-2F
   DHCP Enabled. . . . . . . . . . . : Yes
   Autoconfiguration Enabled . . . . : Yes
   Link-local IPv6 Address . . . . . : fe80::523:d516:15da:d7c8%10(Preferred)
   IPv4 Address. . . . . . . . . . . : 155.69.142.106(Preferred)
   Subnet Mask . . . . . . . . . . . : 255.255.252.0
   Lease Obtained. . . . . . . . . . : Wednesday, December 18, 2013 11:05:31 AM
   Lease Expires . . . . . . . . . . : Tuesday, January 07, 2014 11:05:29 PM
   Default Gateway . . . . . . . . . : 155.69.143.254
   DHCP Server . . . . . . . . . . . : 155.69.3.8
   DHCPv6 IAID . . . . . . . . . . . : 234890513
   DHCPv6 Client DUID. . . . . . . . : 00-01-00-01-19-F2-A6-8C-00-25-11-89-50-2F

   DNS Servers . . . . . . . . . . . : 155.69.3.8
                                       155.69.3.9
   Primary WINS Server . . . . . . . : 155.69.4.83
   Secondary WINS Server . . . . . . : 155.69.5.54
   NetBIOS over Tcpip. . . . . . . . : Enabled
```

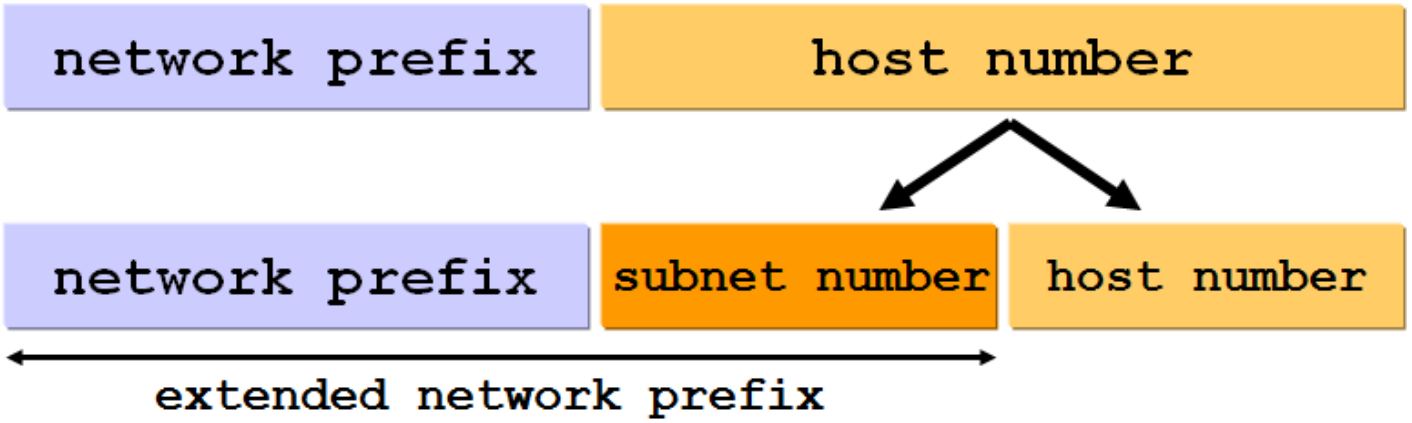
Subnetting
(RFC 950)

- Why? For easier management or other reasons, an organization may wish to partition its single large network into smaller networks called subnets.



Subnetting
(RFC 950)

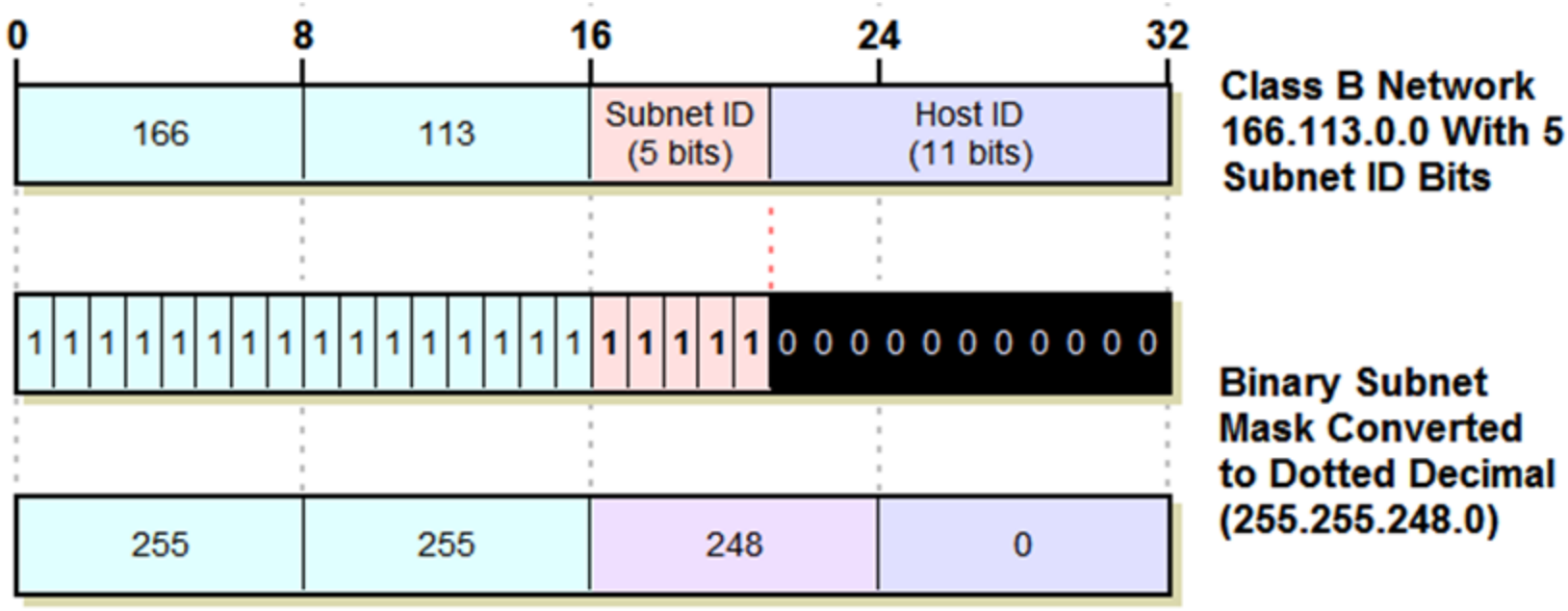
- How? Add another level of hierarchy to the IP address structure.



  - Subnets are only visible within the organization.
  - Hence, organization is free to decide the number of bits for subnet and host numbers.
  - Externally, the organization network is still viewed as a single large network.

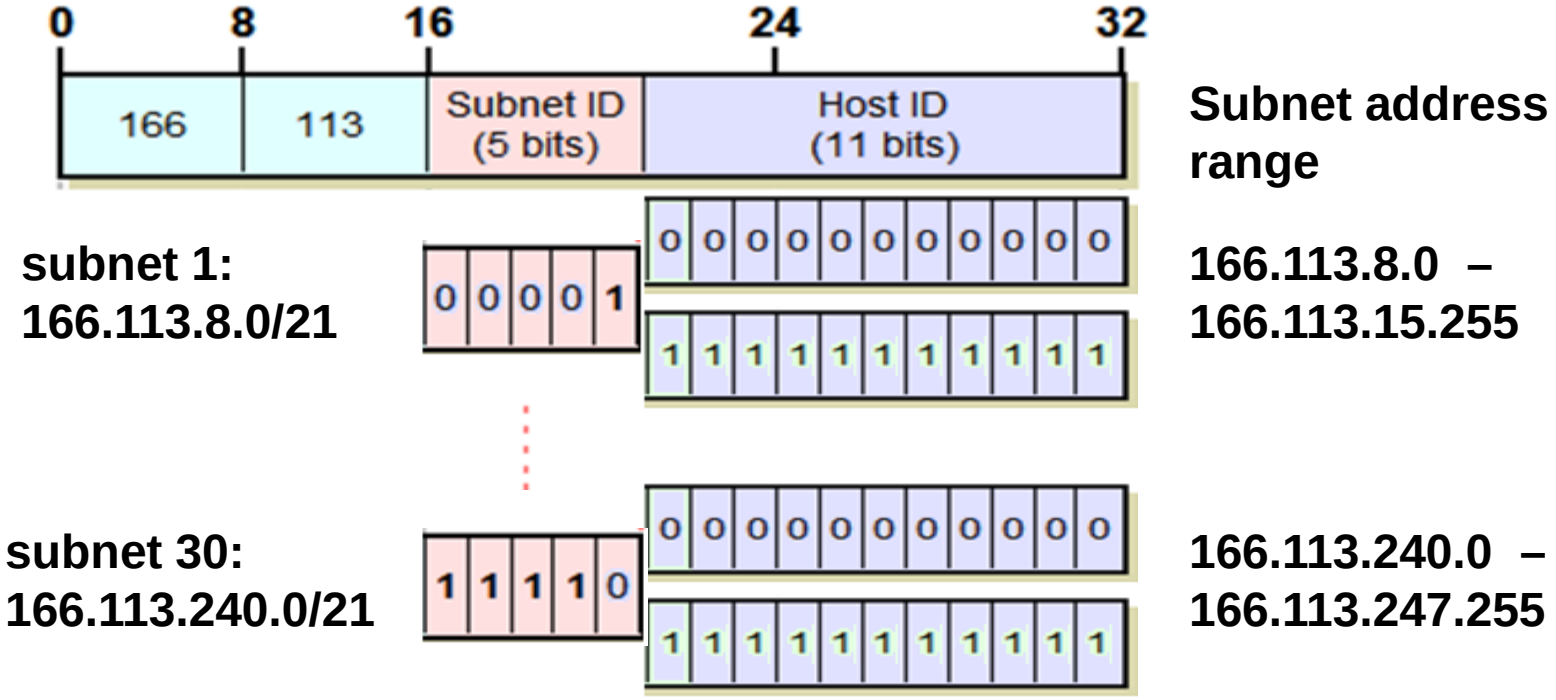
Subnet Masks

- **To indicate the length of extended network prefix, use a subnet mask w.x.y.z (bits corresponding to extended network prefix are set to ‘1’s, and ‘0’s otherwise.)**



Subnet Address Calculation

- 'slash' notation: a.b.c.d/x where x indicates # bits for extended network prefix



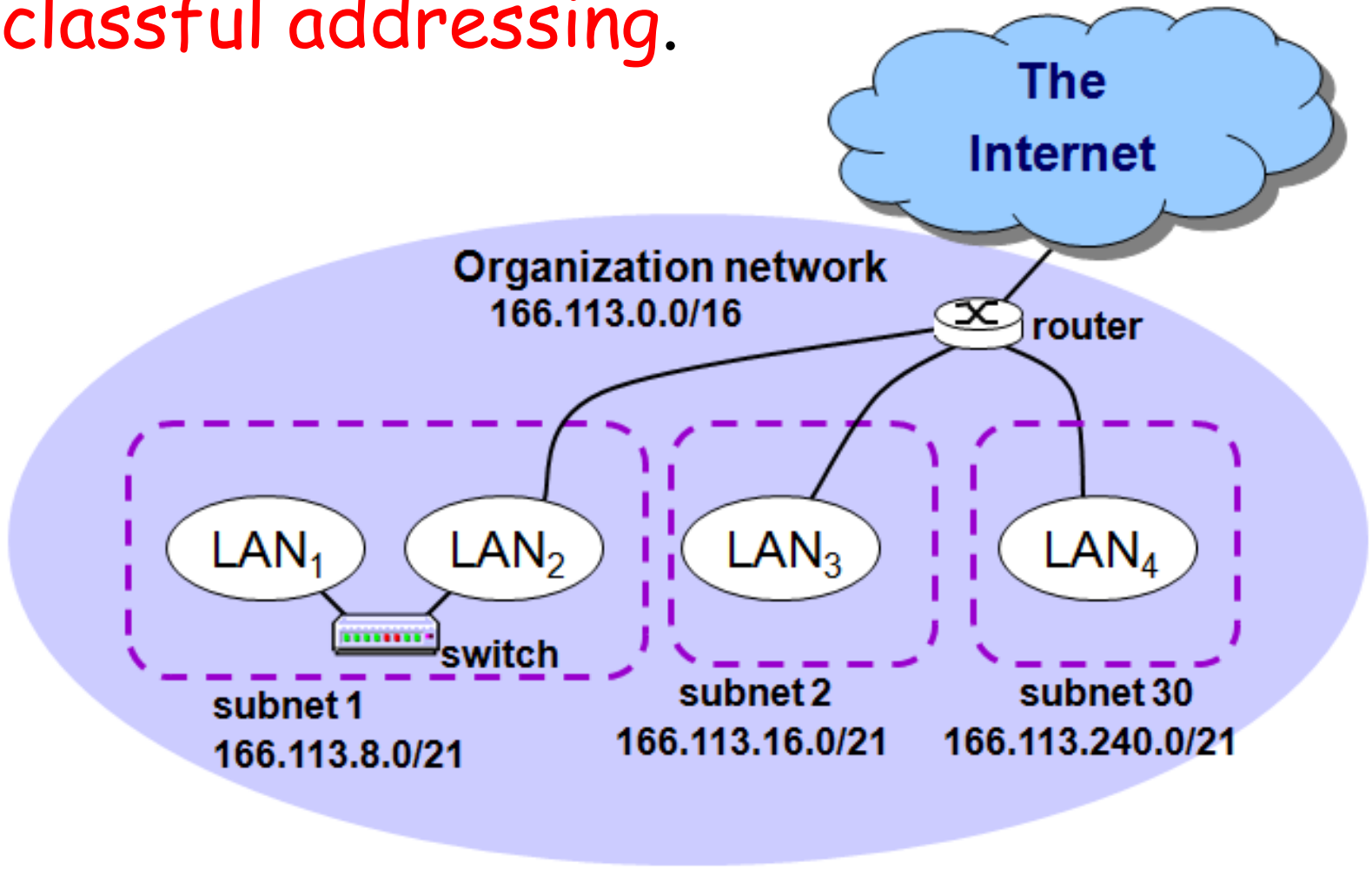
Maximum # hosts in each subnet = $2^{11} - 2 = 2046$ because

- host ID = all '0's indicates network/subnet ID number
- host ID = all '1's indicates broadcast address

Subnet Broadcasting

- **Subnet broadcast (set host address to all 1s):**
  - **Say Class B address: 166.113.x.x**
  - **Subnet mask: 255.255.248.0 (ie. /21)**
  - **Then 166.113.15.255 means broadcast to subnet 1**
- **All subnets broadcast (set subnet & host addresses to all 1s):**
  - **166.113.255.255 means broadcast to all hosts in all subnets**

An example subnetting with traditional classful addressing.



Note: Originally, subnet id = all '0's or all '1's are not allowed. But it is possible now with CIDR – to be discussed later.

Are the Destination and Host IP address in the same subnet ?

- **To determine if the destination IP address is within the same subnet**
  - **Exclusive OR the Source IP and Destination IP network address.**
  - **IF Result =0 THEN same subnet ELSE different subnet**

A	B	EX-OR Output
0	0	0
0	1	1
1	0	1
1	1	0

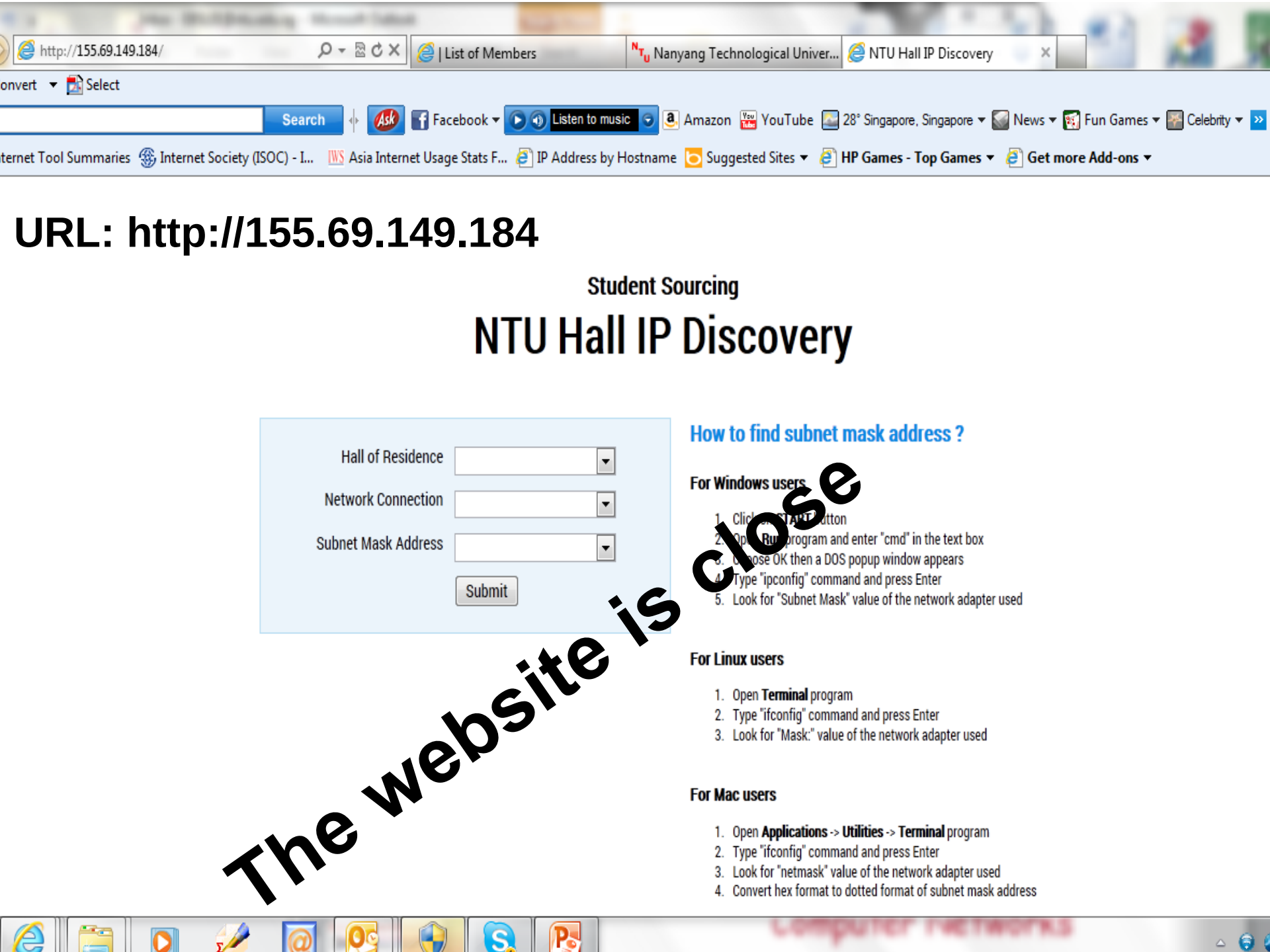
Subnet

If((Host IP) ^(Destination IP) & SubnetMask) = 0)
THEN
{
ARP to get Destination MAC address
Send packet to Destination IP
}
ELSE
{
Send packet to designated Router
}

Where “^” is the bitwise EXOR and
“&” bitwise AND function

IP network address assignment

- **What is the IP address allocated to my organisation ?**
  - **Calculate the total number of hosts in our network, eg. 155.69.0.0/20 means we have a maximum of 2^{12} - 2 number of host.**
- **How many dept/school/division are there in the organisation ?**
  - **This will determine the number of subnets. May need additional subnets for backbone.**
- **How many host in each dept/school/division ?**
  - **This would determine the subnet mask for each subnet. Always have buffer.**
- **How do I assign the address within a subnet ?**
  - **Assign servers IP address from the lowest to the highest host address.**
  - **Assign client IP address from highest to lowest host address numbers.**


http://155.69.149.184/
List of Members
Nanyang Technological Univer...
NTU Hall IP Discovery
Search
Facebook
Listen to music
Amazon
YouTube
28° Singapore, Singapore
News
Fun Games
Celebrity
Internet Tool Summaries
Internet Society (ISOC) - I...
Asia Internet Usage Stats F...
IP Address by Hostname
Suggested Sites
HP Games - Top Games
Get more Add-ons
URL: http://155.69.149.184
Student Sourcing
NTU Hall IP Discovery
Hall of Residence
Network Connection
Subnet Mask Address
Submit
How to find subnet mask address ?
For Windows users
Type "ipconfig" command and press Enter
Look for "Subnet Mask" value of the network adapter used
For Linux users
Open Terminal program
Type "ifconfig" command and press Enter
Look for "Mask:" value of the network adapter used
For Mac users
Open Applications -> Utilities -> Terminal program
Type "ifconfig" command and press Enter
Look for "netmask" value of the network adapter used
Convert hex format to dotted format of subnet mask address
The website is close

Records found: 25

Hall of Residence	Network connection	IP Address	Subnet Mask
6	Wired	155.69.140.17	255.255.255.248
7	Wired	172.20.72.99	255.255.252.0
7	Wired	172.20.72.99	255.255.252.0
13	Wired	172.20.133.16	255.255.252.0
14	Wired	172.20.145.50	255.255.252.0
5	Wired	172.20.53.77	255.255.252.0
12	Wired	172.20.125.120	255.255.252.0
8	Wireless	172.20.84.217	255.255.255.0
13	Wireless	172.20.132.106	255.255.255.0
13	Wired	172.20.132.25	255.255.255.255
4	Wireless	172.20.45.2	255.255.255.0
3	Wired	172.20.174.64	255.255.255.0
14	Wired	172.20.146.251	255.255.252.0
7	Wired	172.20.73.32	255.255.252.0
4	Wired	172.20.46.60	255.255.252.0
3	Wired	172.20.173.119	255.255.255.0
12	Wireless	172.20.124.226	255.255.255.0
12	Wired	172.20.125.141	255.255.255.0
8	Wired	172.20.84.96	255.255.252.0
1	Wireless	172.20.12.119	255.255.255.0
15	Wired	172.20.153.19	255.255.252.0

How are IP addresses subnetted in the Halls ?

- **IP address range : 172.20.X.X**
- **Subnet mask : 255.255.252.0 (/22)**

- **Halls allocation (z is just a digit from 0-9)**
  - **172.20.1z.* Hall 1**
  - **172.20.7z.* Hall 7**
  - **172.20.13z.* Hall 13**
  - **172.20.15z.* Hall 15**

IP address efficiency

- **Example: Hall 13**
  - **IP address : 172.20.133.16/22**
  - **Subnet mask : 255.255.252.0**
  - **Range of subnet IP addresses :**
    - **172.20.132.0 till 172.20.135.255**

Did not use 172.20.$z.X where z is 0 or 1, eg. 172.20.130.0 till 172.20.131.255

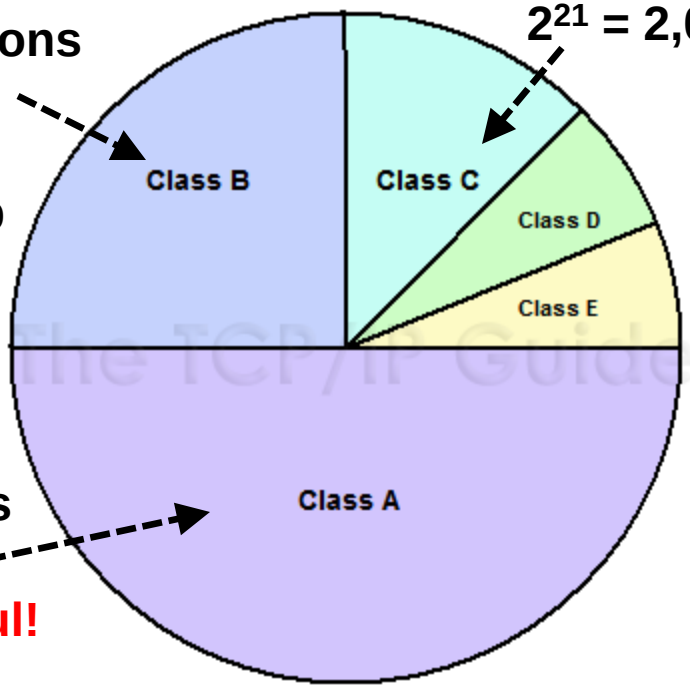
3rd byte(Binary)		Decimal
100000	10	130
100000	11	131
100001	00	132
100001	01	133
100001	10	134
100001	11	135
100010	00	136

IPv4 Address Exhaustion!!!

IP classful addressing problems:

- **Inefficient use of address space**

2^{14} = 16,384 organizations owns 25% of total IP addresses, with each Class B supporting up to 65,534 hosts. Do they need that many?



2^{21} = 2,097,152 organizations owns 25% of total IP addresses. But each Class C supporting only 254 hosts is too small for many.

128-2 = 126 organizations owns 50% of total IP addresses – very wasteful!

- **32-bit IP address is just not enough for today's global network and future growth.**

ARIN
American Registry for Internet Numbers

Your IPv4 address is 155.69.14.241

SEARCH Whois
advanced search

NUMBER RESOURCES | PARTICIPATE | POLICIES | FEES & INVOICES | KNOWLEDGE | ABOUT US

ARIN ONLINE

Username and password are case sensitive.

username: *new user?*

password: *assistance*

log in

About ARIN Online

REGIONAL INTERNET REGISTRIES

SEARCH THIS SECTION

Advanced Search

Regional Internet Registries (RIRs) are nonprofit corporations that administer and register Internet Protocol (IP) address space and Autonomous System (AS) numbers within a defined region. RIRs also work together on joint projects.

REGIONS

- All
- AFRINIC
- APNIC
- ARIN
- LACNIC
- RIPE NCC

Registry	Geographic Region
AFRINIC	Africa, portions of the Indian Ocean
APNIC	Portions of Asia, portions of Oceania
ARIN	Canada, many Caribbean and North Atlantic islands, and the United States
LACNIC	Latin America, portions of the Caribbean
RIPE NCC	Europe, the Middle East, Central Asia

IPv4 address exhaustion

- **"Exhaustion" is defined here as the time when the pool of available addresses in each RIR reaches the "last /8 threshold" of 16,777,216 addresses.**

http://www.potaroo.net/tools/ipv4/index.html

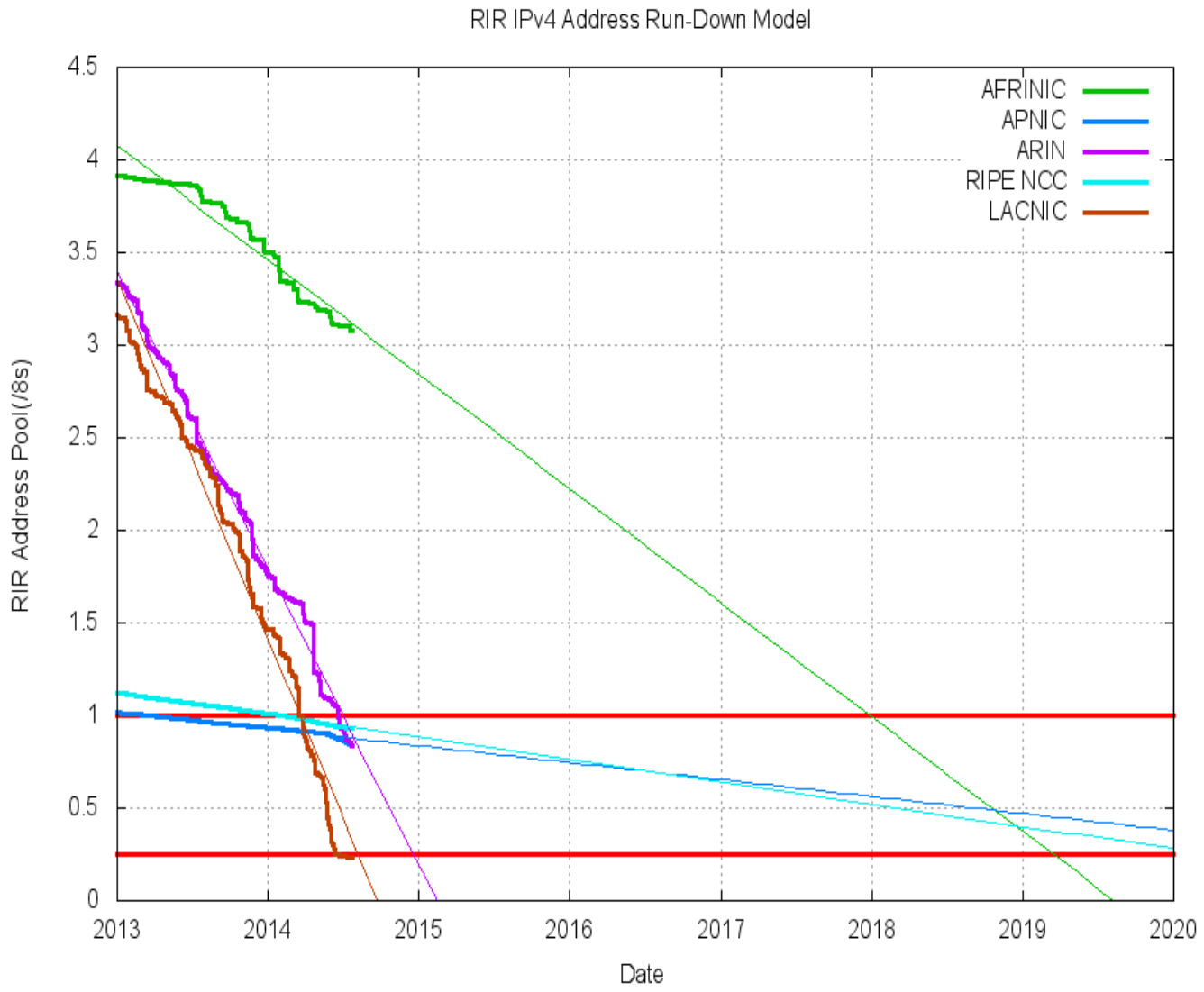
RIR	Projected exhaustion date	Remaining addresses in RIR pool (/8)
APNIC	19 April 2011(actual)	0.8394
RIPE NCC	14 Sept2012(actual)	0.9281
ARIN	19 Feb 2015	0.8340
LACNIC	10 June 2014(actual)	0.2338
AFRINIC	17 July 2019	3.0749

IANA exhausted 3 Feb 2011

Copied from website on 26 Jul 2014

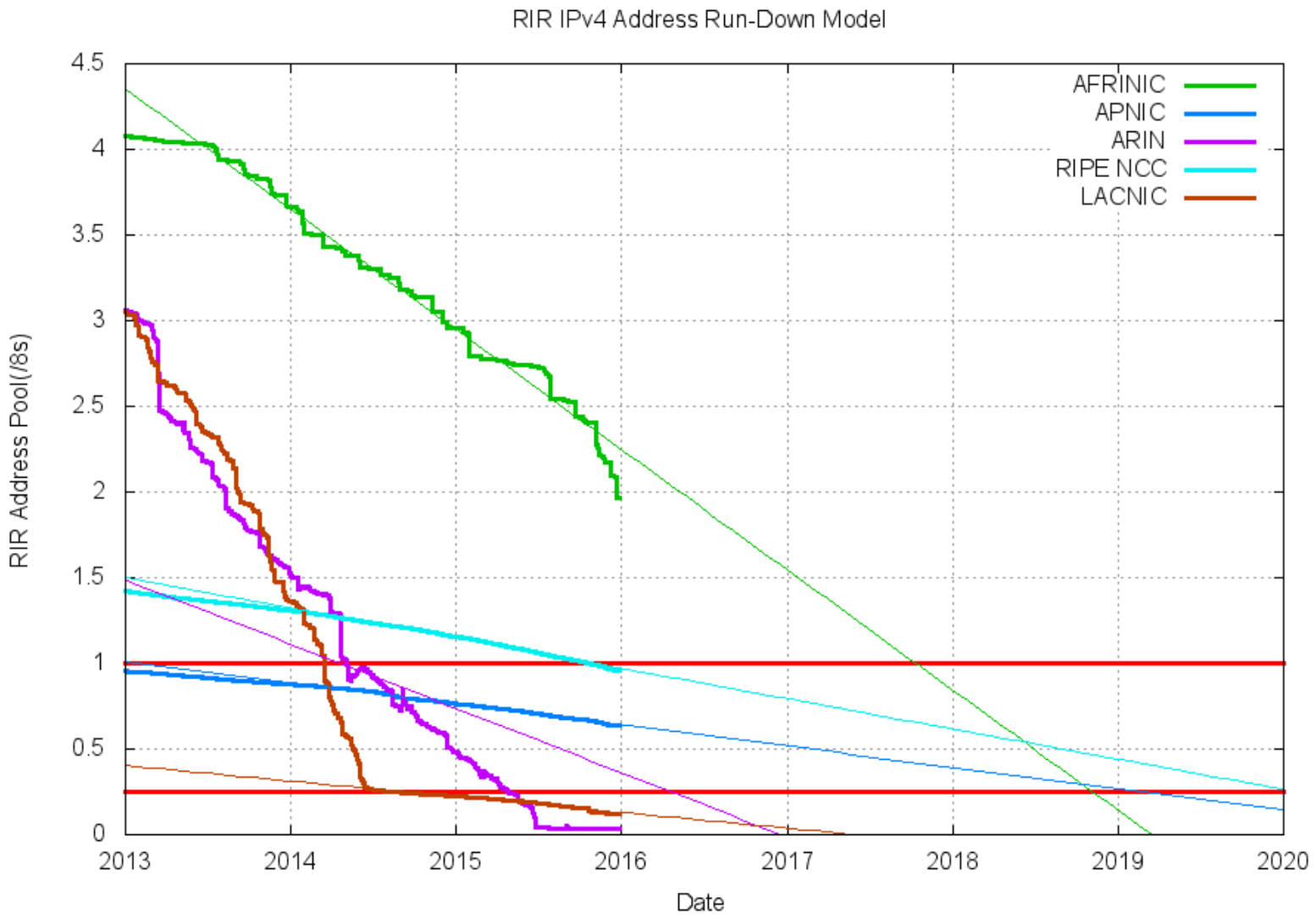
http://www.potaroo.net/tools/ipv4/index.html

(as at 26 July 2014)



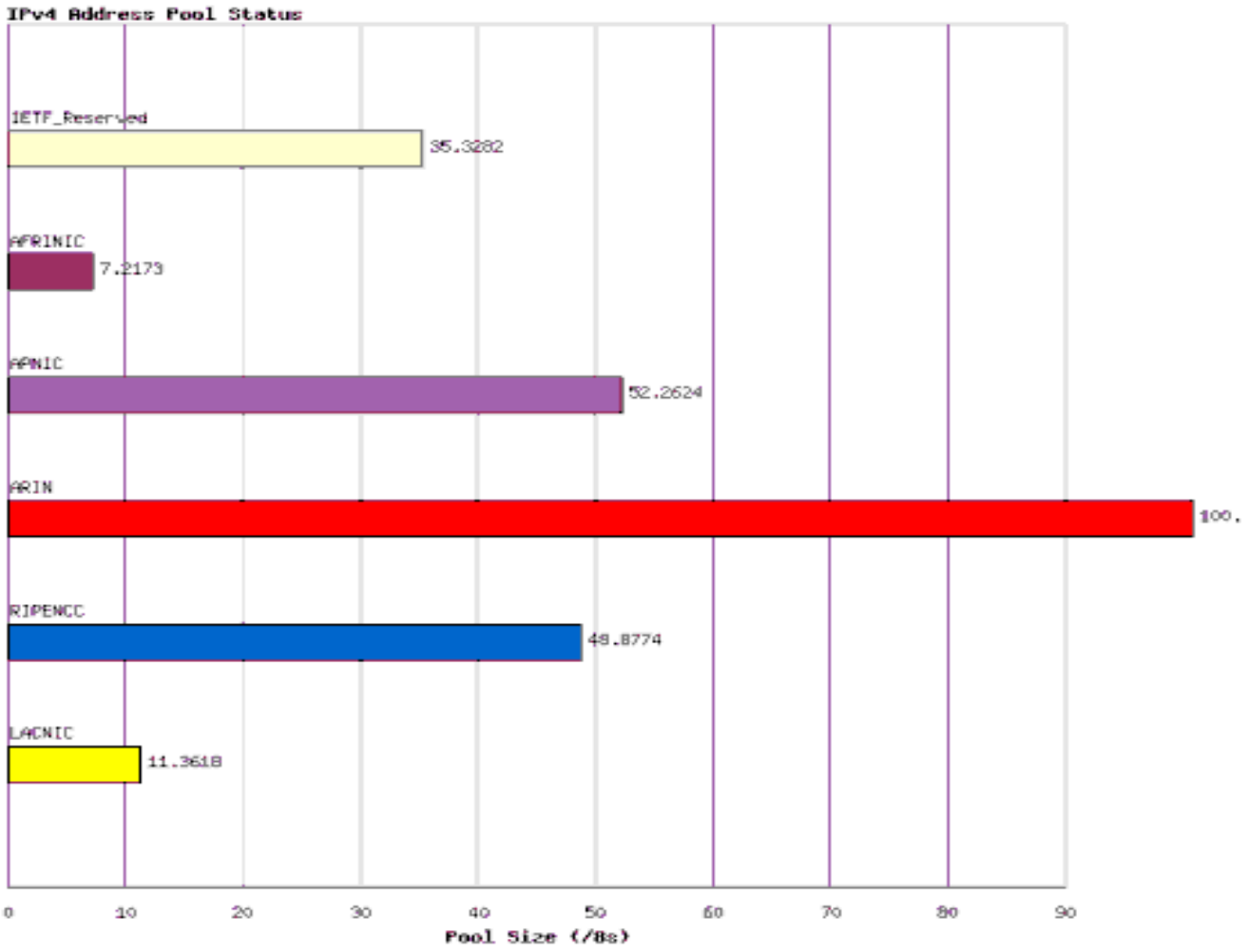
http://www.potaroo.net/tools/ipv4/index.html

(as at 30 Dec 2015)



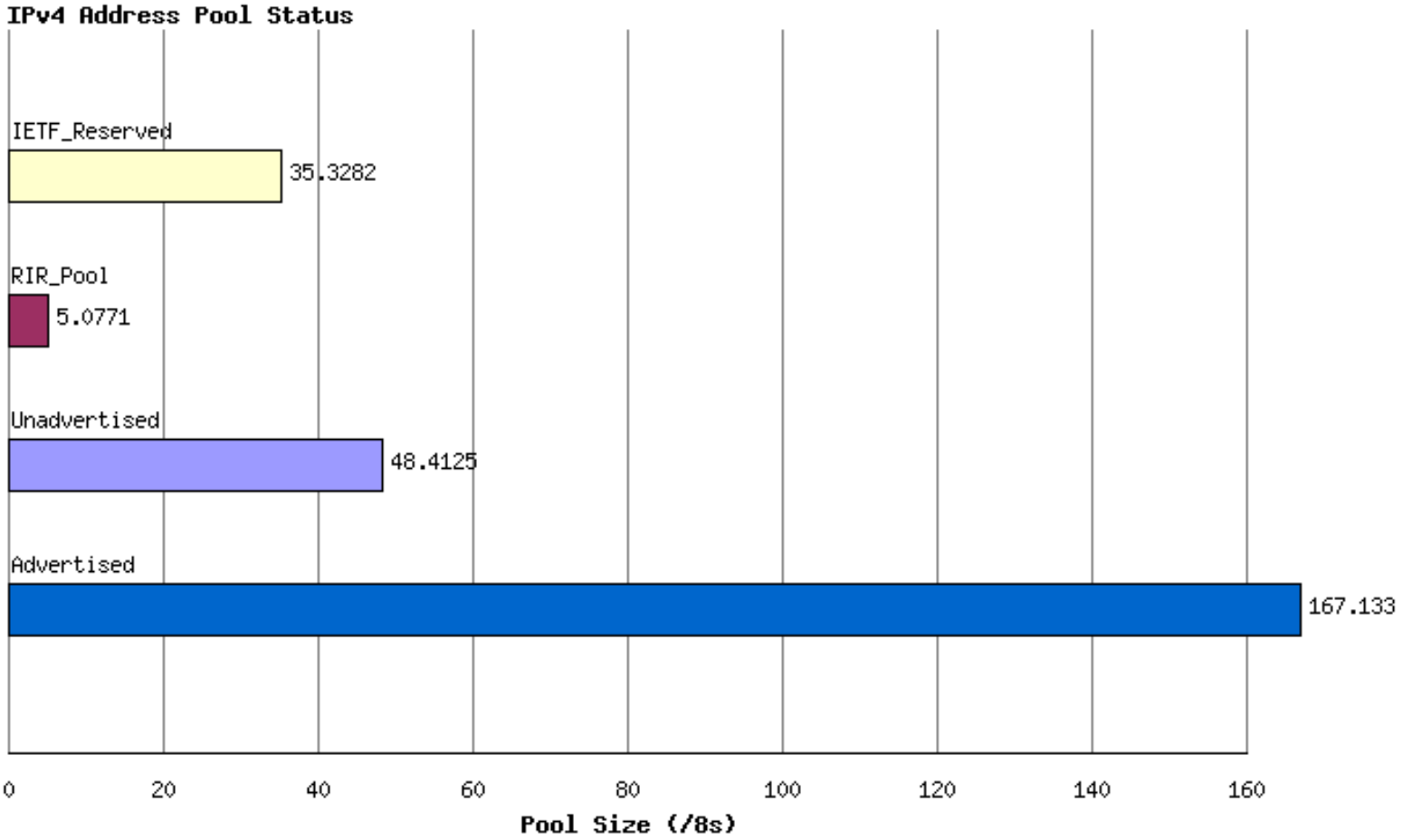
IPv4 address usage

(30 Dec 2015)



IPv4 address usage

(30 Dec 2015)

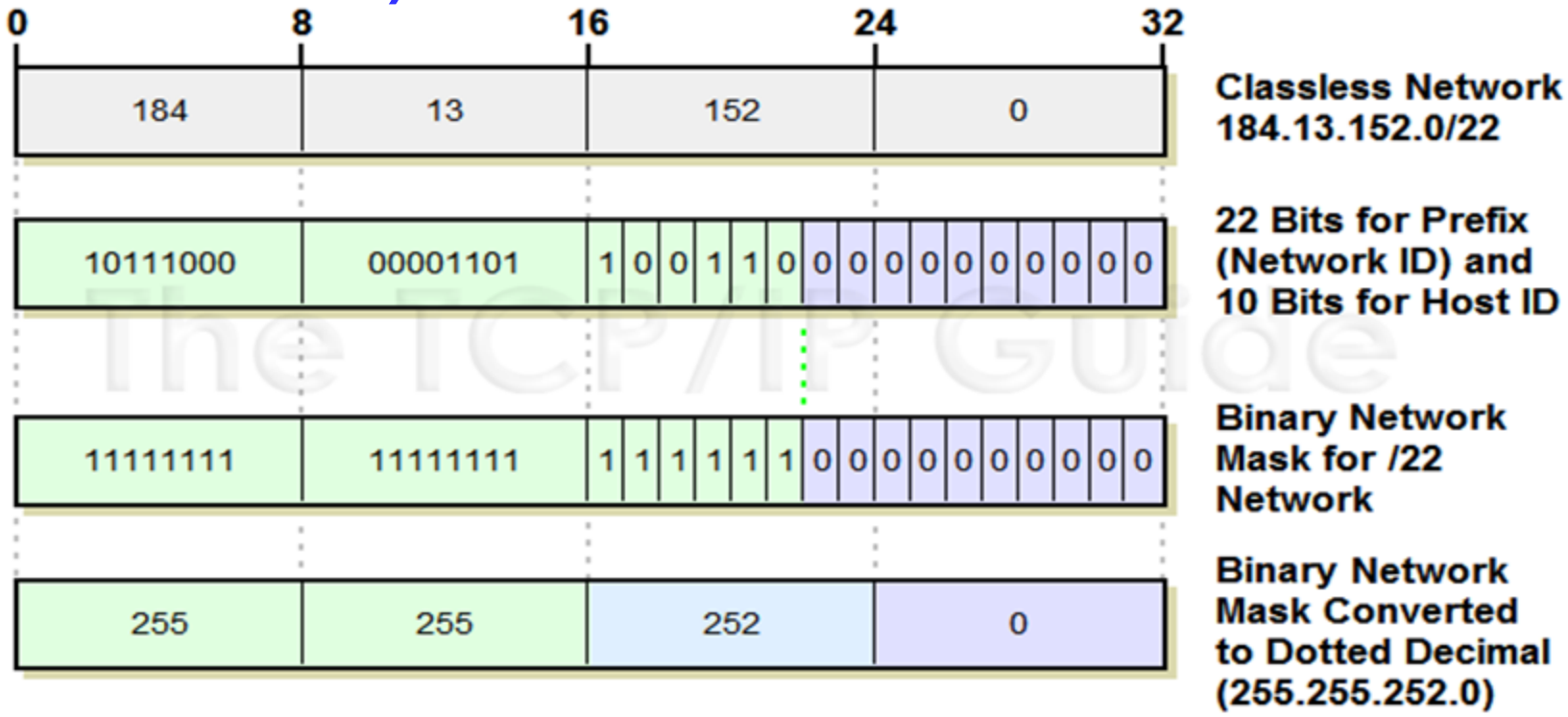


IPv4 Address Exhaustion

- **Solutions to IPv4 Address Exhaustion:**
  - **Classless InterDomain Routing (short-term solution)**
    - **This reduces the wastage in address allocation. Organizations will be given adequate but not excessive address space.**
  - **Network Address Translation (NAT) using Private IP Addresses (will ease but not solve the problem)**
    - **A single machine (usually a router) with an IP address representing many computers behind it. IP addresses require translation.**
  - **IP version 6, 128-bit space (long-term solution)**
    - **It will be large enough to install several billion computers on every square meter of the Earth's surface!**
    - **Problem: People has no motivation to upgrade, so how?**

Classless Inter-Domain Routing
(RFC 1517 - 1519)

- Abandon the notion of classful addressing
- Key concept: length of network id (prefix) can be any length
- Consequence: add a network mask w.x.y.z (similar concept as subnet mask)



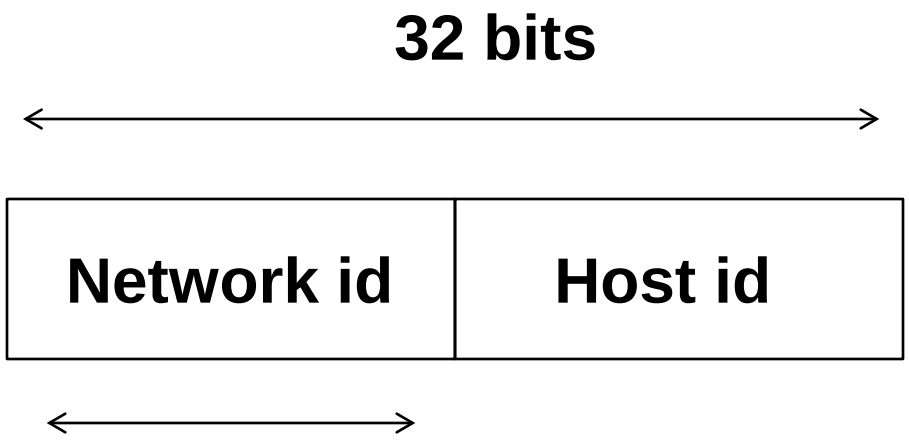
The TCP/IP Guide

CIDR- Subnet Mask

CIDR	Subnet Mask
/8	255.0.0.0
/9	255.128.0.0
/10	255.192.0.0
/11	255.224.0.0
/12	255.240.0.0
/13	255.248.0.0
/14	255.252.0.0
/15	255.254.0.0
/16	255.255.0.0

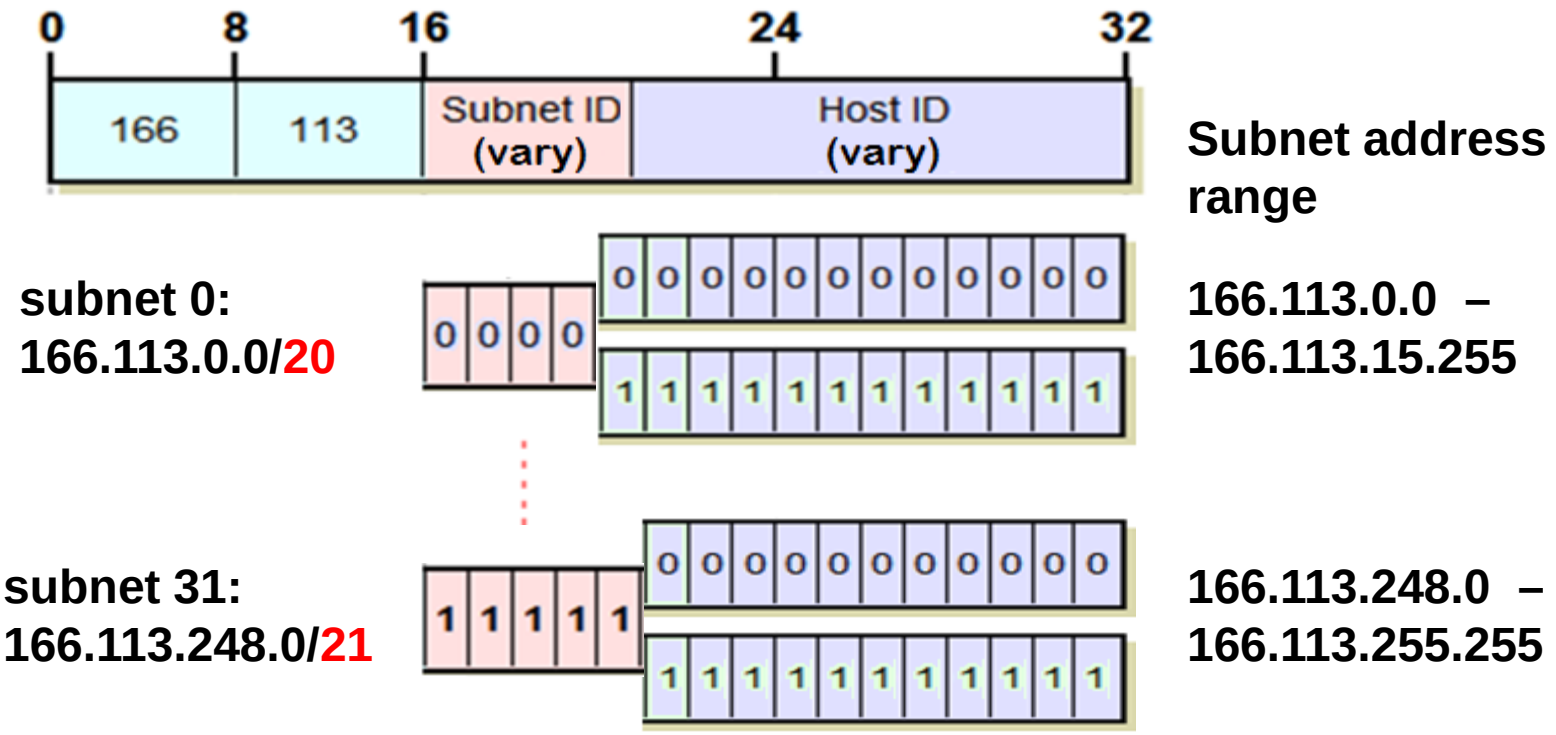
CIDR

- **The IP address consists of the Network Id, and Host id.**
- **“/n” where “n” is the number of bits allocated to Network id.**



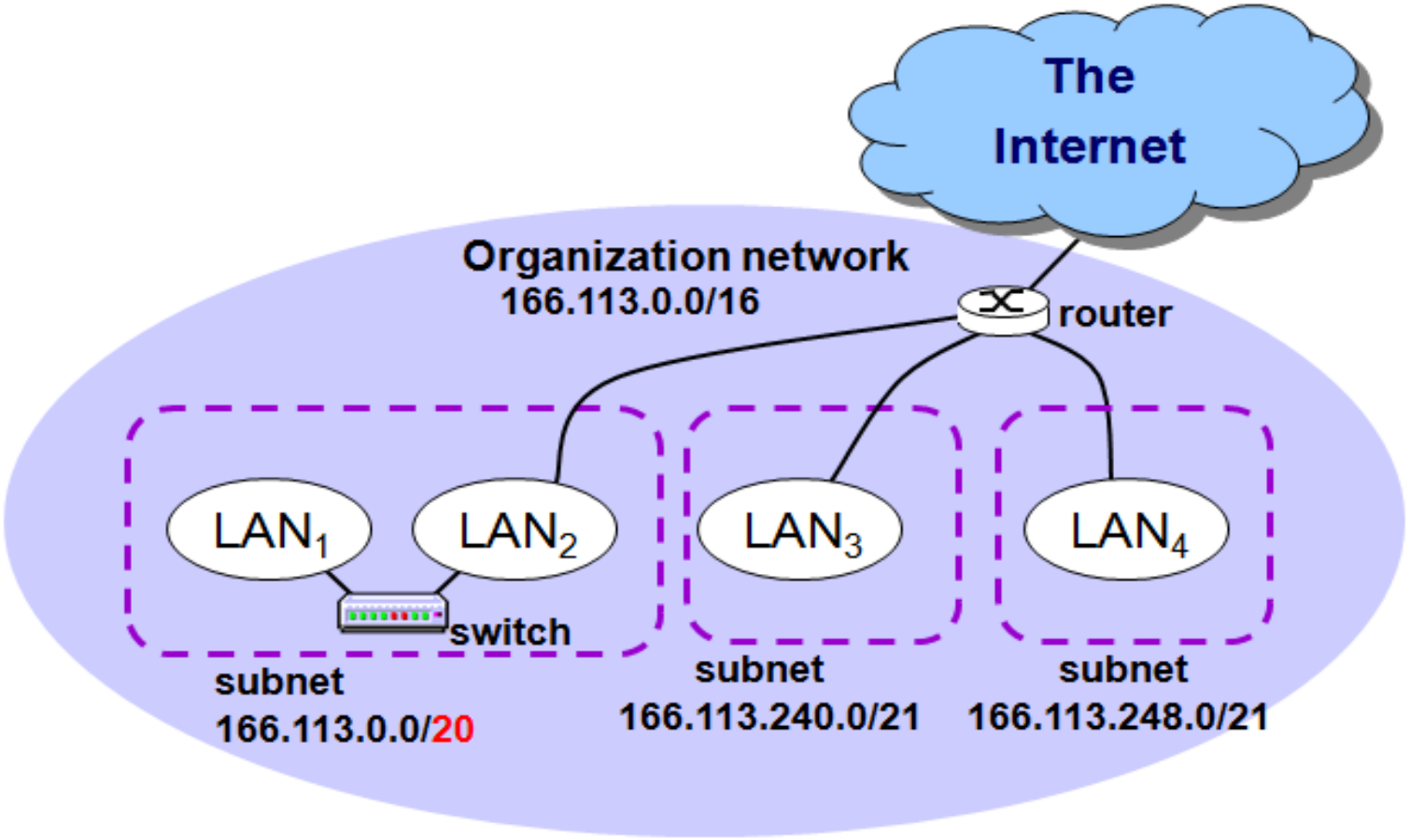
W.X.Y.Z /n

In fact, CIDR can be extended to subnetting. Hence, it's now possible to have variable-length subnet masks.



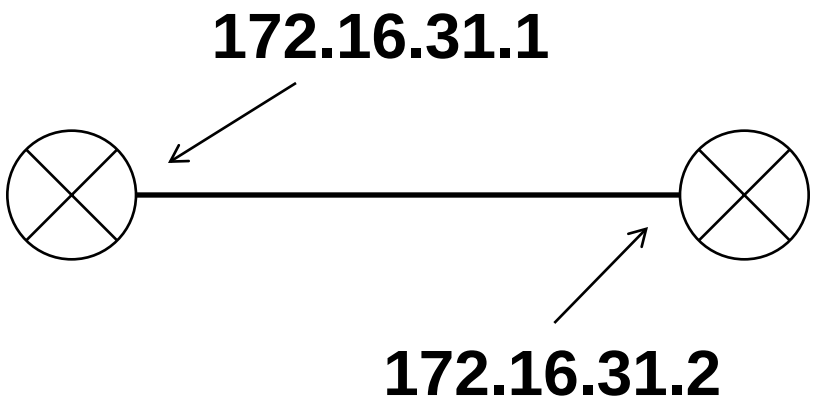
Note: The idea of all subnets broadcast is made obsolete. Therefore, subnet numbers including all '0's and all '1's can now be used.

An example subnetting with CIDR.



Router to router link

- **An IP address needs to be assigned to each active interface of a router.**
- **To optimised the use of IP address we usually assigned a /30 to the link**



APRICOT 2015: Elise Gerich looking back at the ABC's of the Number Registry System

By George Michaelson on 2 Mar 2015

Category: Tech matters

Tags: addressing, APRICOT2015, IANA, ICANN, IP, RIRs, routing

Like Share 0 Tweet 0

Blog home

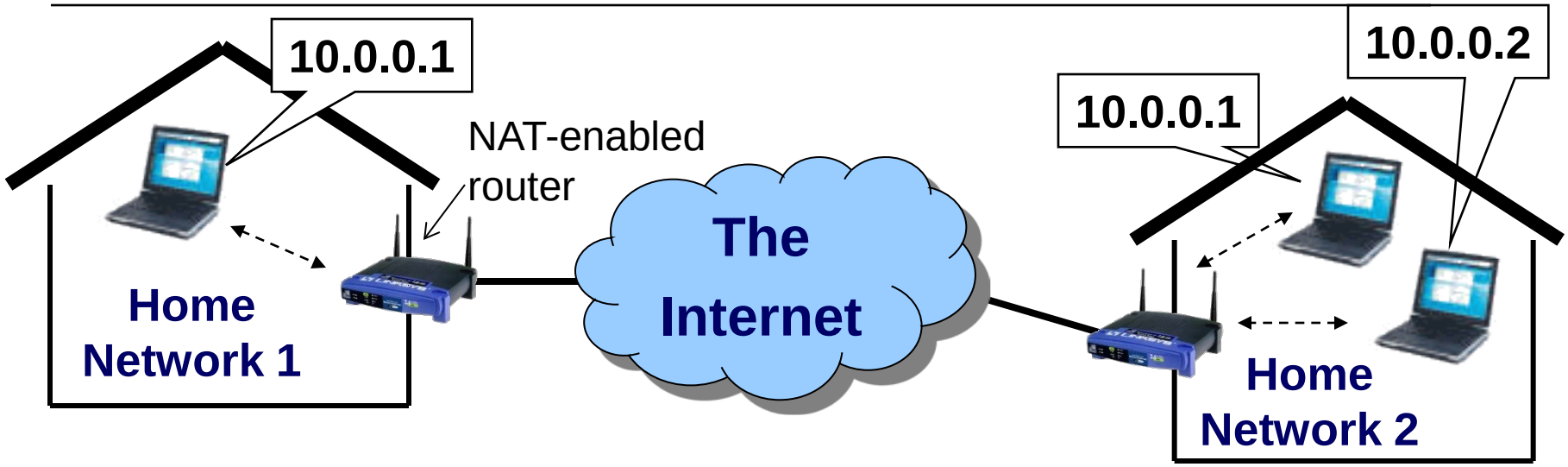
In her APRICOT 2015 keynote, Elise Gerich from ICANN presented a historical review of the Internet numbering system that led us to where we stand today. Elise is uniquely suited to talk to this subject, having a 22 year history of leadership in the Internet, and her early role in the NSFNet function. Elise provided much of the critical infrastructure which led directly from an Academic/Research network, to the fully commercially viable Internet we have now. Elise did this for MERIT, the agency which had management oversight of the emerging network.

https://blog.apnic.net/2015/03/02/apricot-2015-elise-gerich-looking-back-at-the-abcs-of-the-num

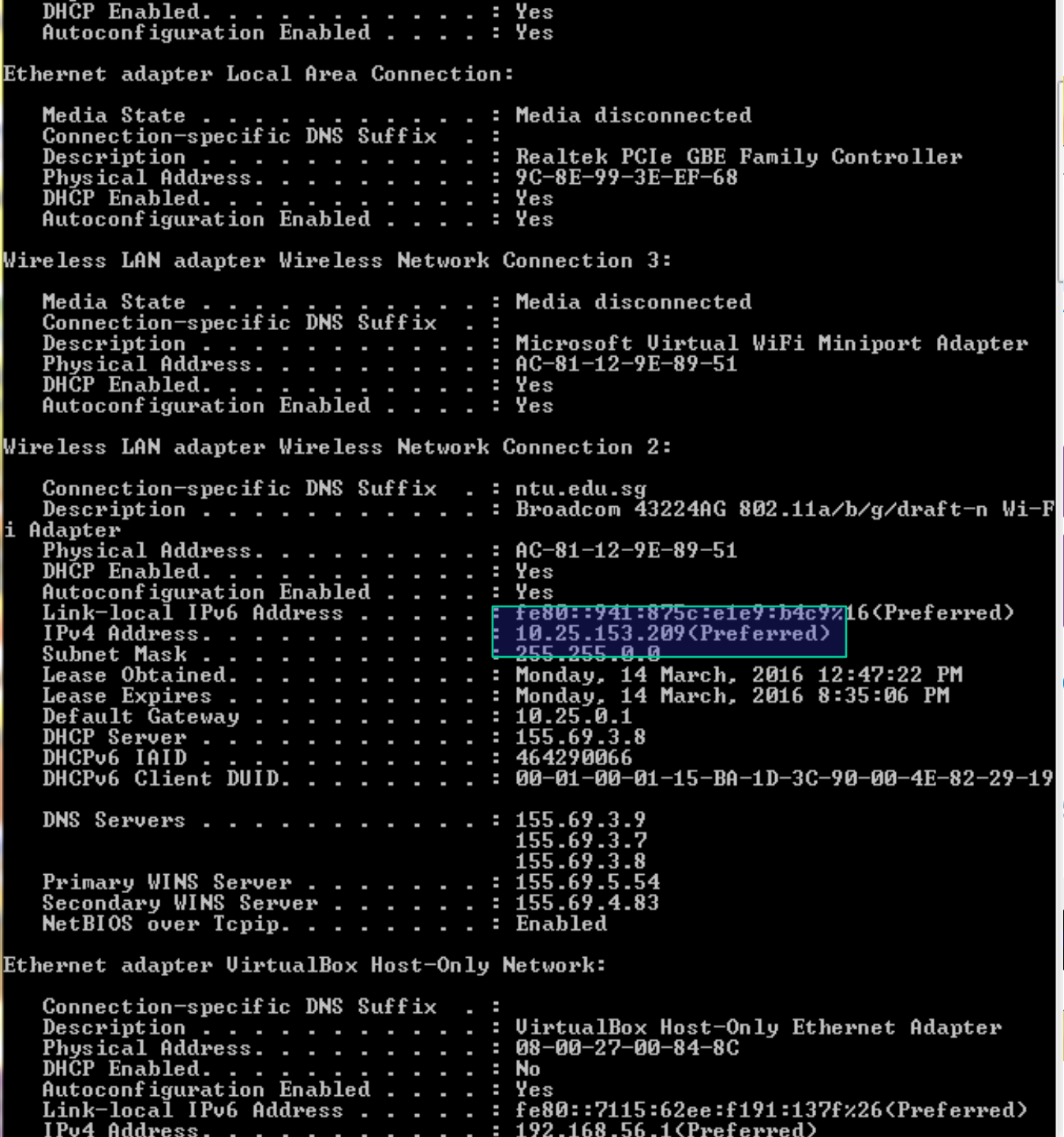
Network Address Translation (NAT)

(RFC 1918) Private IP address for Private Internet:

- **10.0.0.0/8 (10.0.0.0 – 10.255.255.255)**
- **172.16.0.0/12 (172.16.0.0 – 172.31.255.255)**
- **192.168.0.0/16 (192.168.0.0 – 192.168.255.255)**



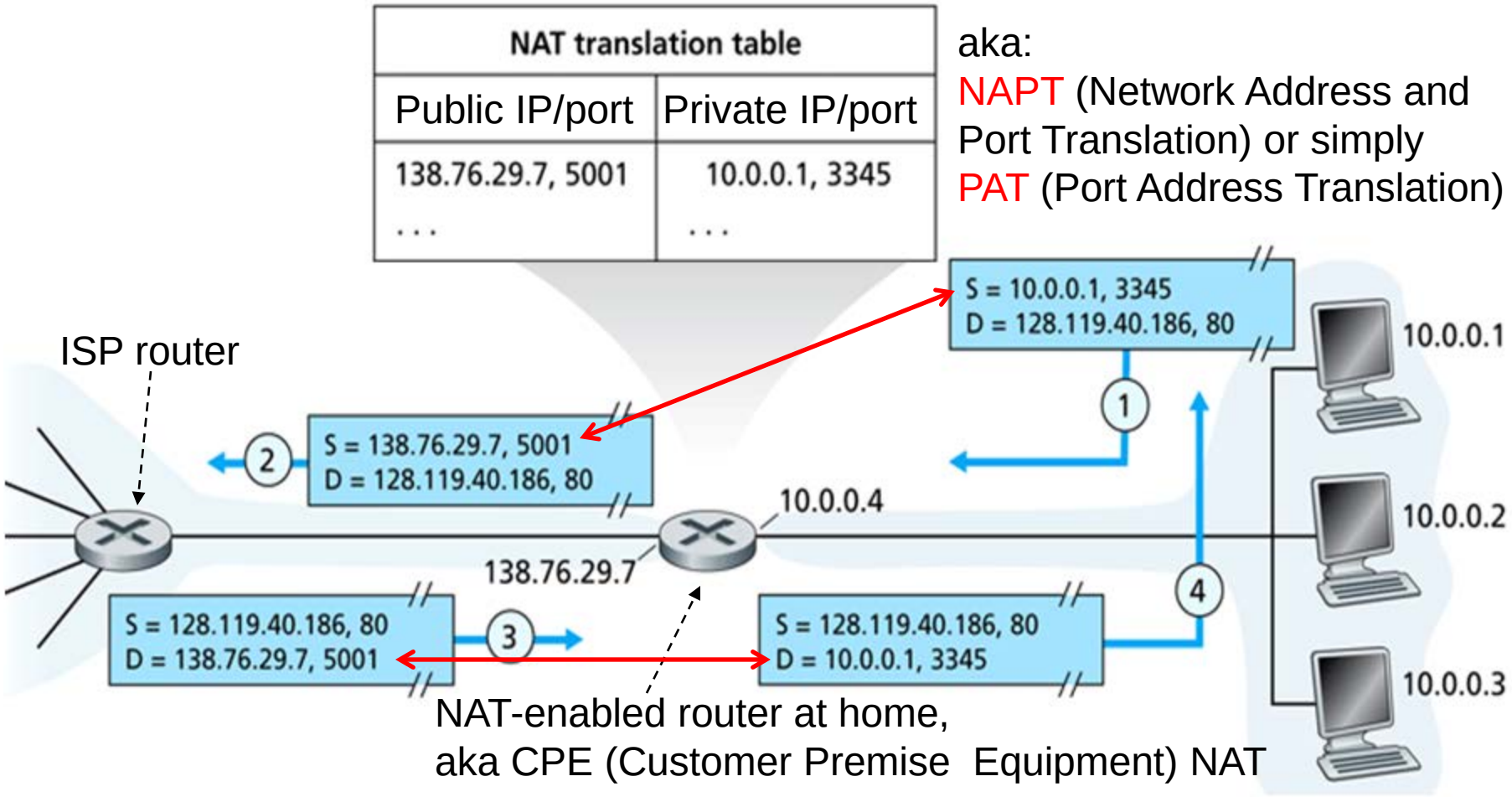
- **Private IP addresses will not be forwarded into the Internet.**
- **Hence, different private networks can re-use the same private IP addresses.**


DHCP Enabled. . . . . : Yes
Autoconfiguration Enabled . . . . : Yes
Ethernet adapter Local Area Connection:
Media State . . . . . : Media disconnected
Connection-specific DNS Suffix . :
Description . . . . . : Realtek PCIe GBE Family Controller
Physical Address. . . . . : 9C-8E-99-3E-EF-68
DHCP Enabled. . . . . : Yes
Autoconfiguration Enabled . . . . : Yes
Wireless LAN adapter Wireless Network Connection 3:
Media State . . . . . : Media disconnected
Connection-specific DNS Suffix . :
Description . . . . . : Microsoft Virtual WiFi Miniport Adapter
Physical Address. . . . . : AC-81-12-9E-89-51
DHCP Enabled. . . . . : Yes
Autoconfiguration Enabled . . . . : Yes
Wireless LAN adapter Wireless Network Connection 2:
Connection-specific DNS Suffix . : ntu.edu.sg
Description . . . . . : Broadcom 43224AG 802.11a/b/g/draft-n Wi-F
i Adapter
Physical Address. . . . . : AC-81-12-9E-89-51
DHCP Enabled. . . . . : Yes
Autoconfiguration Enabled . . . . : Yes
Link-local IPv6 Address . . . . . : fe80::941:875c:ele9:b4c9%16(Preferred)
IPv4 Address. . . . . : 10.25.153.209(Preferred)
Subnet Mask . . . . . : 255.255.0.0
Lease Obtained. . . . . : Monday, 14 March, 2016 12:47:22 PM
Lease Expires . . . . . : Monday, 14 March, 2016 8:35:06 PM
Default Gateway . . . . . : 10.25.0.1
DHCP Server . . . . . : 155.69.3.8
DHCPv6 IAID . . . . . : 464290066
DHCPv6 Client DUID. . . . . : 00-01-00-01-15-BA-1D-3C-90-00-4E-82-29-19
DNS Servers . . . . . : 155.69.3.9
155.69.3.7
155.69.3.8
Primary WINS Server . . . . . : 155.69.5.54
Secondary WINS Server . . . . . : 155.69.4.83
NetBIOS over Tcpip. . . . . : Enabled
Ethernet adapter VirtualBox Host-Only Network:
Connection-specific DNS Suffix . :
Description . . . . . : VirtualBox Host-Only Ethernet Adapter
Physical Address. . . . . : 08-00-27-00-84-8C
DHCP Enabled. . . . . : No
Autoconfiguration Enabled . . . . : Yes
Link-local IPv6 Address . . . . . : fe80::7115:62ee:f191:137f%26(Preferred)
IPv4 Address. . . . . : 192.168.56.1(Preferred)

rch
Community
Blog
About
Your IP address: 155.69.191.254
ATE
WHAT IS APNIC?
Feedback
sses for the Asia Pacific
light
1:34 PM
14/3/2016

Network Address Translation
(RFC 2663, 3022)

By using a NAT-enabled router, only 1 IP address is required from ISP to support the whole private network to connect to Internet.

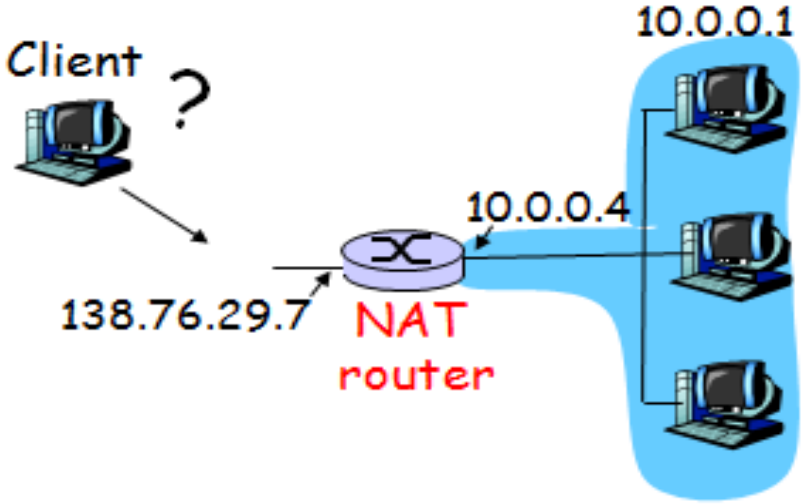


NAT Traversal Problem

However, NAT is controversial:

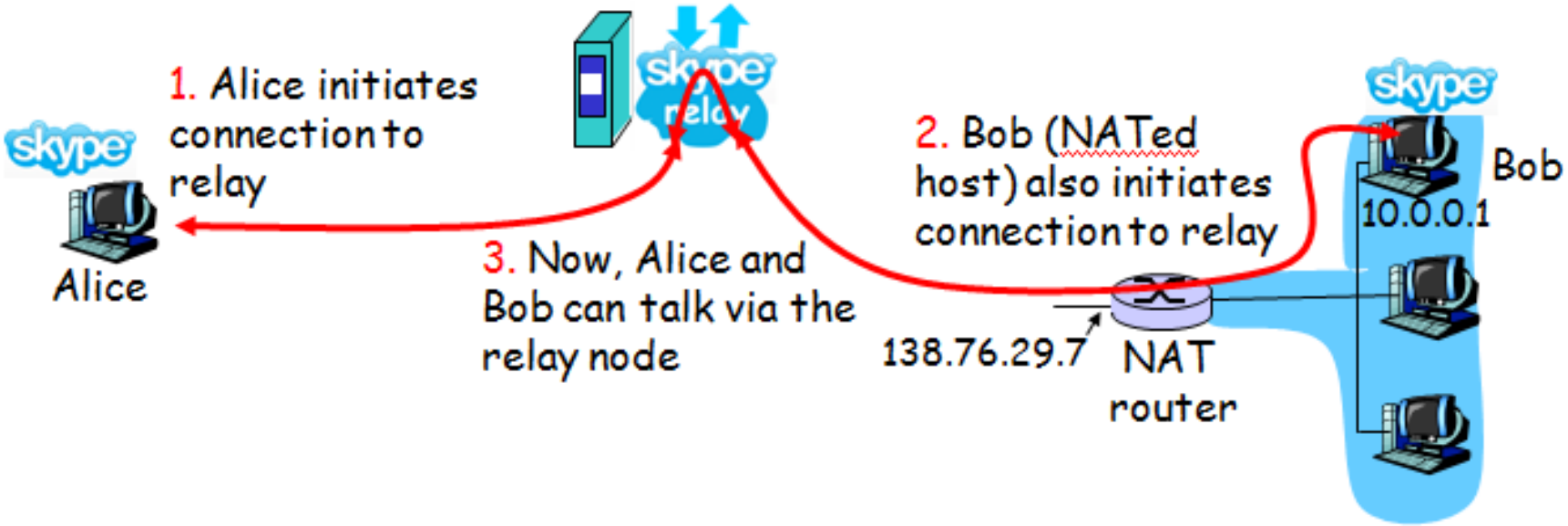
- Routers (layer-3 device) should only process up to layer-3 (NAT-enabled routers process up to transport layer to change port number).
- Violates host-to-host concept of layer-3
  - As a result, application developers, especially P2P, need to consider the possibility of a host behind a NAT

How to contact a host behind a NAT (also called NATed host) since its IP address (e.g. 10.0.0.1) is not recognizable in the Internet?



To overcome NAT traversal problem, a possible workaround is to use a relay like Skype.

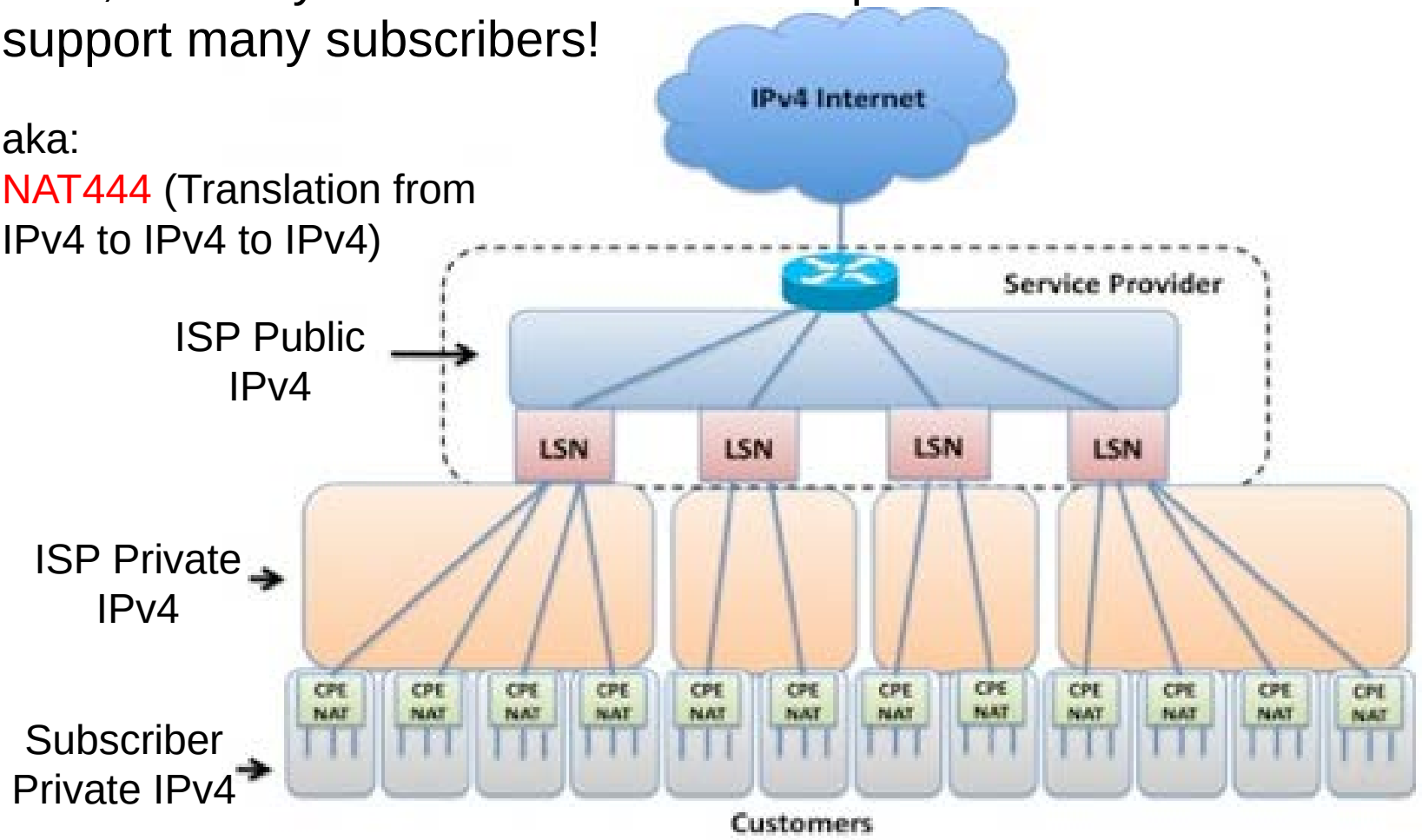
- Alice and Bob sign in and remain connected with their assigned non-NATed super peers.
- To call Bob, Alice informs her super peer, which will inform Bob's super peer, which will in turn inform Bob.
- When Bob accepts the call, their super peers will inform them to connect to a non-NATed relay node.



To further conserve IPv4 addresses, ISP can similarly implement Carrier-Grade NAT (CGN) / Large-Scale NAT (LSN).

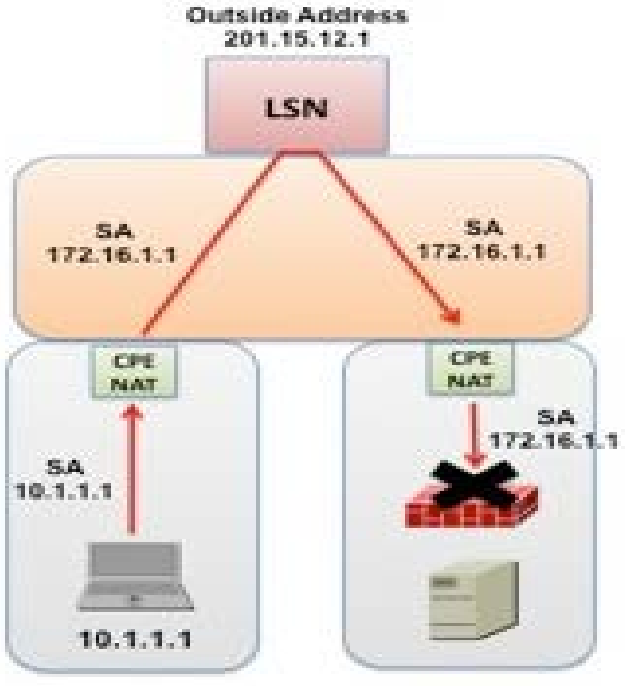
Now, ISP only needs a small block of public IP addresses to support many subscribers!

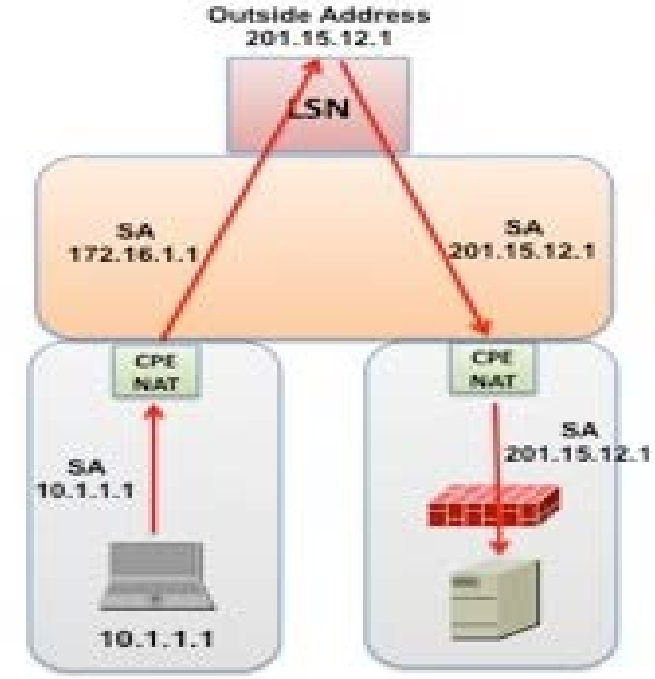
aka:
NAT444 (Translation from IPv4 to IPv4 to IPv4)



However, there are potential issues with Carrier-Grade NAT (CGN) / Large-Scale NAT (LSN):

- Routing problem – if subscribers' private IP addresses are the same as ISP's private IP addresses
- Delay – from one subscriber to another subscriber within same ISP network, still needs to go through NAT since most routers/firewalls will block incoming packets with source address SA = private IP



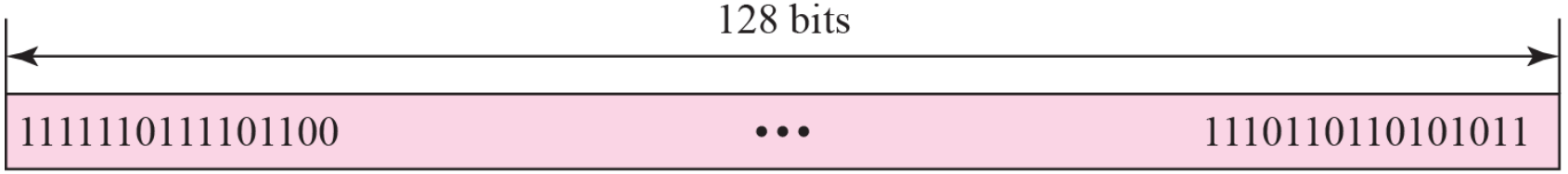


IPv6 (version 6)

Standardised since 1995 (RFC 1883, 2460)

Main enhancement in IPv6:

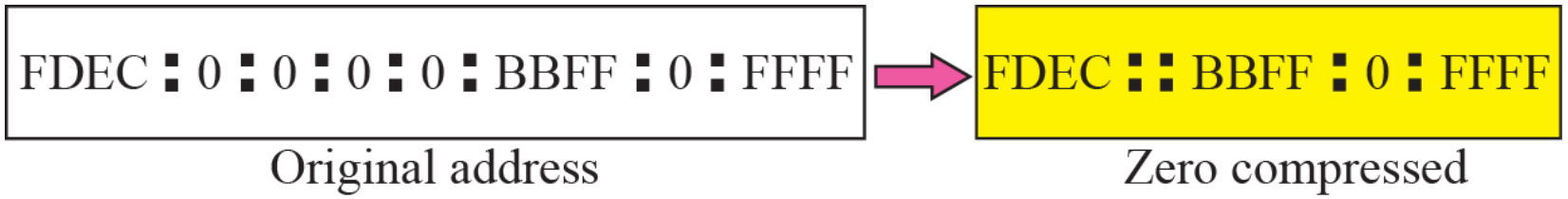
- **Expanded Address Space: 128-bit IPv6 address**



- **Colon hexadecimal notation:**

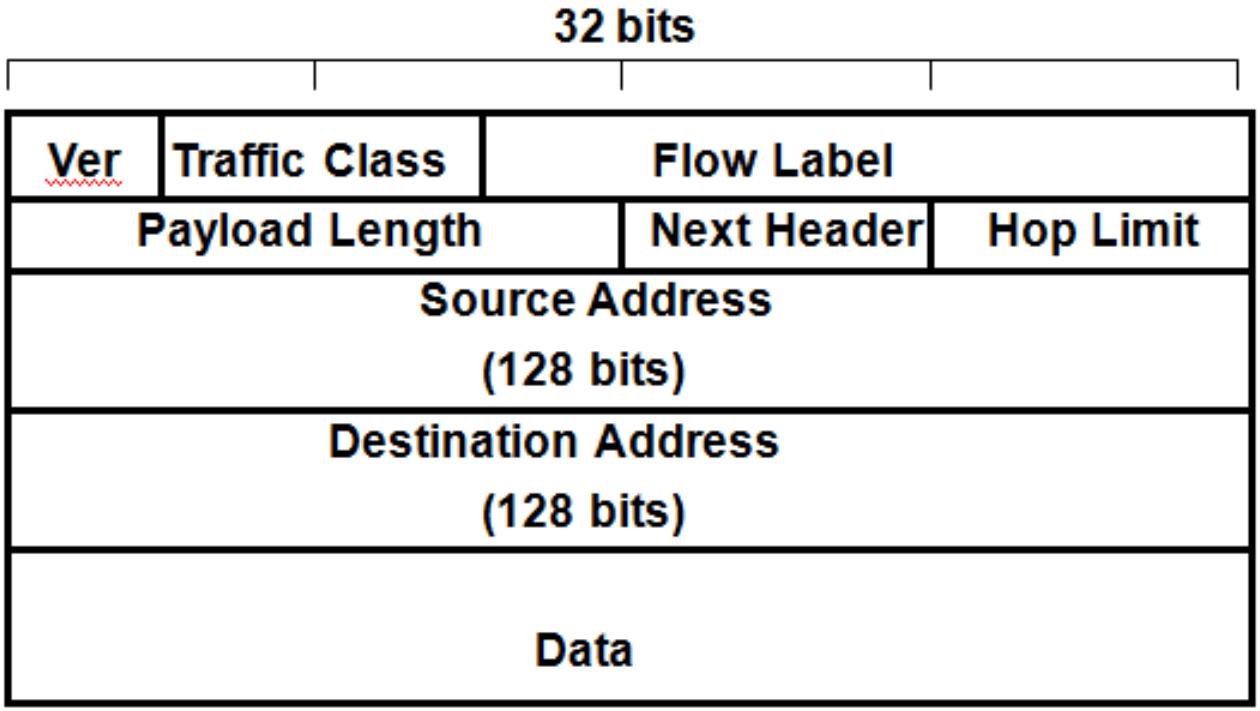
FDEC : BA98 : 7654 : 3210 : ADBF : BBFF : 2922 : FFFF

- **Abbreviated notation:**
  - **within each 16-bit value, 0000 can be written as 0**
  - **consecutive groups of 0s can be replaced by ::**



IPv6 Header

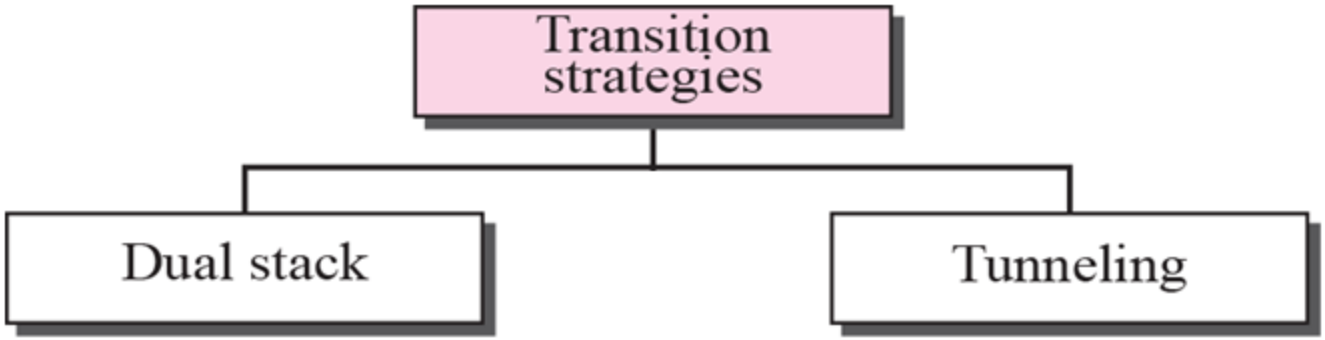
- **Simplification of Header: faster processing.**



Note: IPv4 and IPv6 are NOT compatible. Only the first 4-bit (ver field) in IPv4 and IPv6 headers are the same to distinguish them.

Transition from IPv4 to IPv6

Not all routers can be upgraded simultaneously. So, how will the network operates with mixed IPv4 and IPv6 routers?



Further reading: RFC 4213

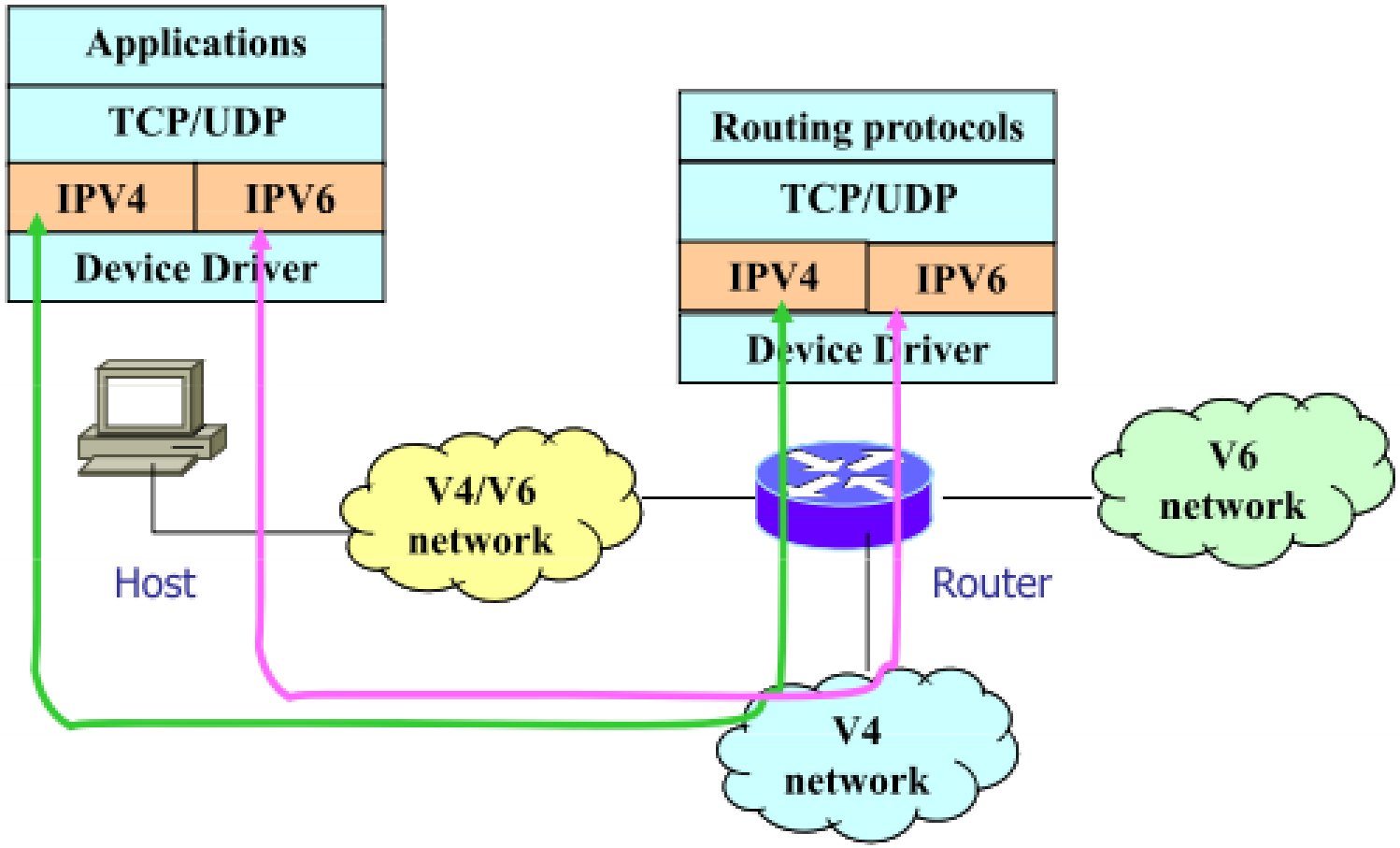
Note on terminology:

Encapsulate – lower layer carrying upper layer data

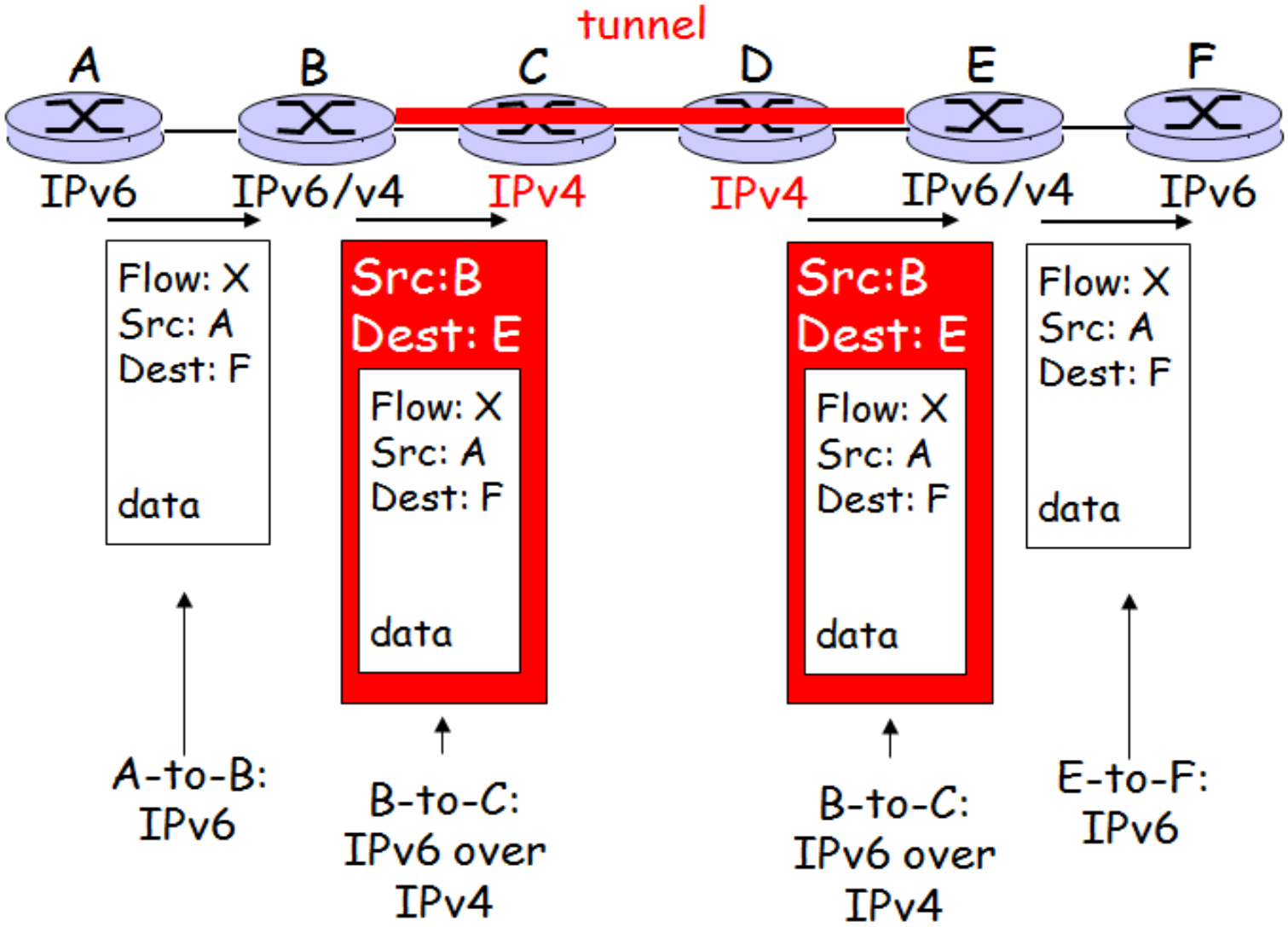
Tunnel – upper layer carrying same or lower layer data

Dual-Stack/Dual-IP Layer

Implement both IPv4 and IPv6 at hosts and routers; e.g. Windows 7 is IPv6 ready.



Tunneling of IPv6 over IPv4

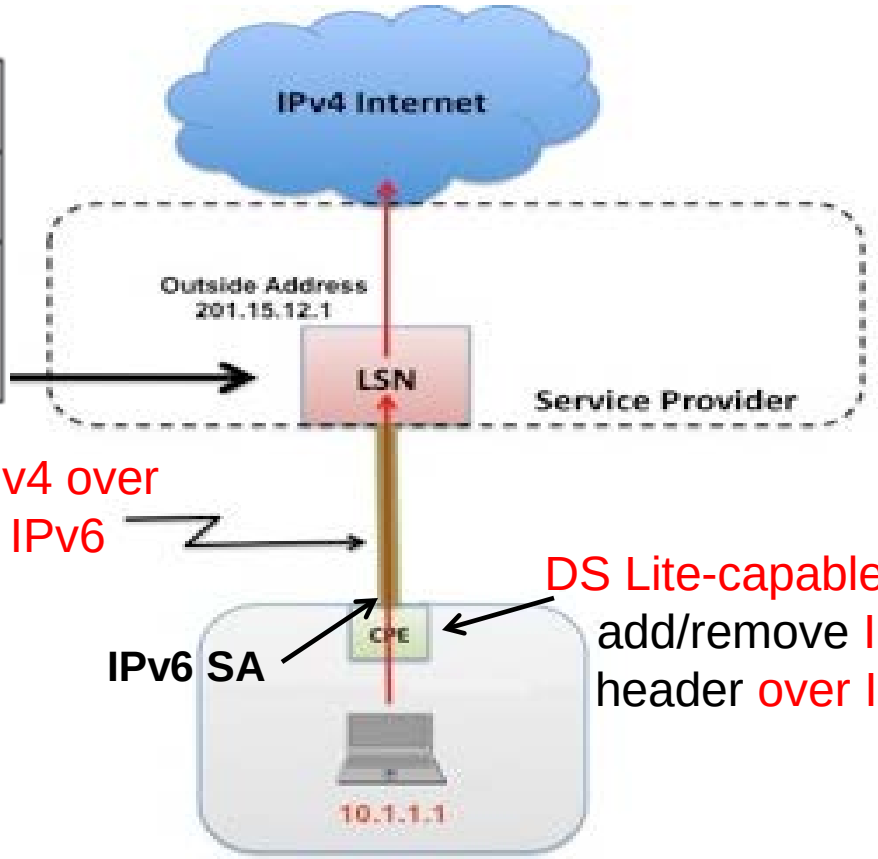


In reality, after more than 10 years, Internet fails to transit to IPv6 and is still mainly IPv4! As a result, Dual-Stack Lite (RFC 6333) is now proposed.

IPv4 addresses are now almost exhausted! So, Dual-Stack Lite (DS Lite) is designed for ISP to share IPv4 addresses among its subscribers by using IPv4 over IPv6 and NAT.

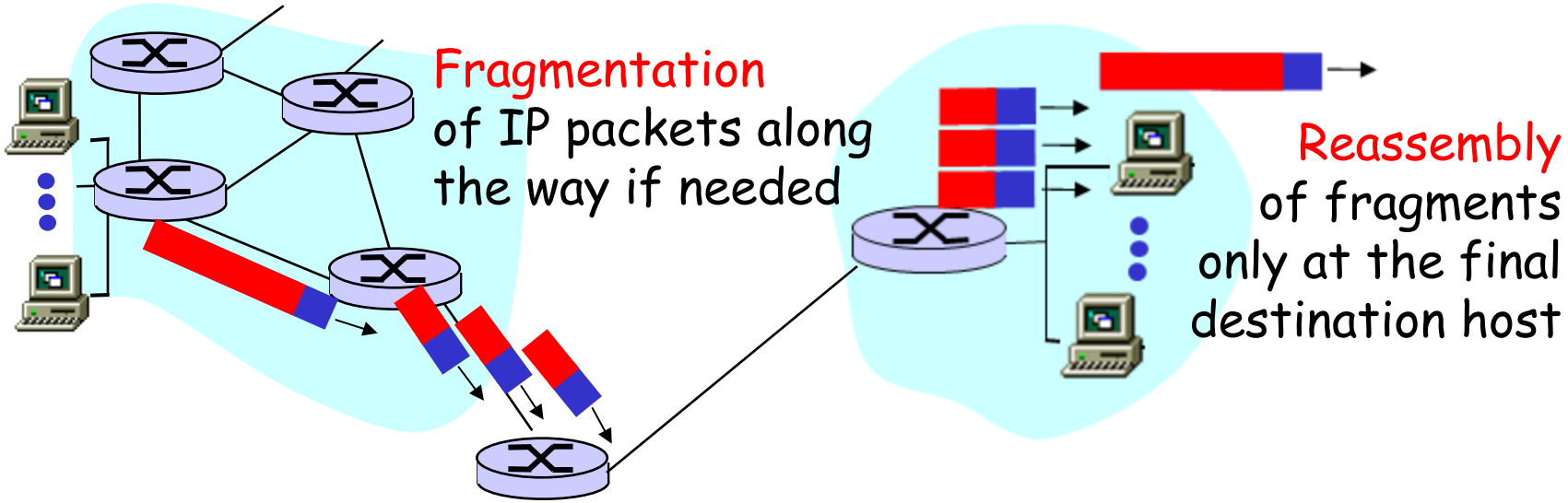
NAT translation table	
Inside IP/port	**Outside IP/port**
IPv6 SA + 10.1.1.1: 3345	**201.15.12.1: 5001**
. . .	. . .

DS Lite-capable LSN: add/remove IPv6 header over IPv4, and perform NAT44

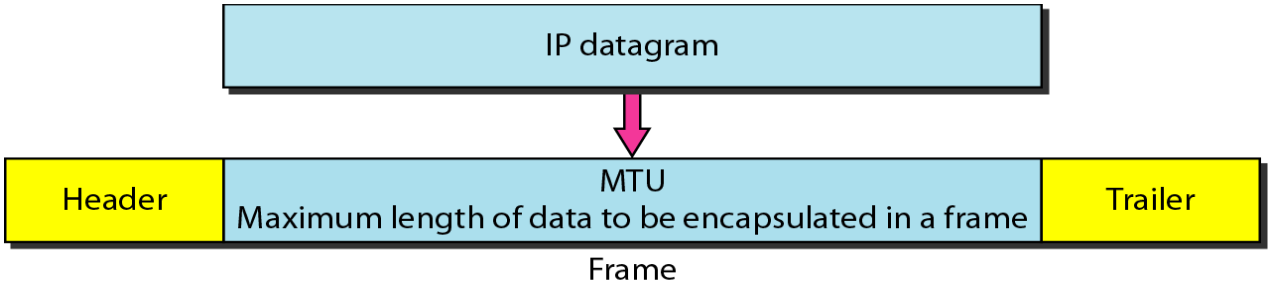


DS Lite-capable CPE: add/remove IPv6 header over IPv4

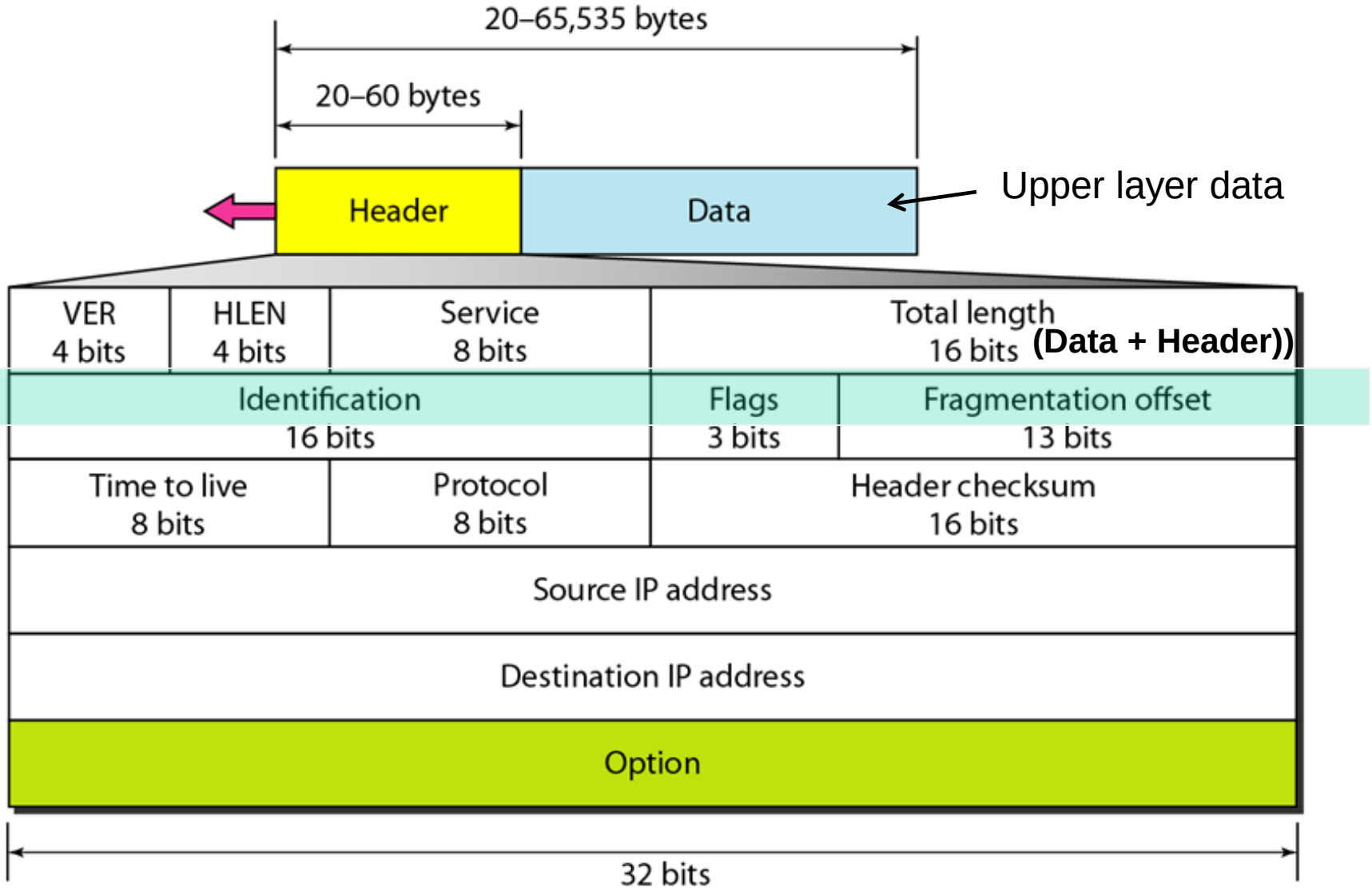
IP Fragmentation & Reassembly
(RFC 791 and 815)



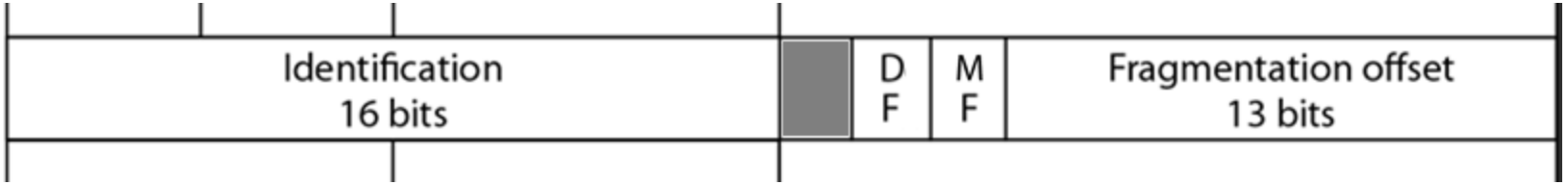
Different networks have different MTU (Maximum Transfer Unit). IP packet size(IP header+data) in each network ≤ MTU of that network.



IPv4 header
(RFC 791)



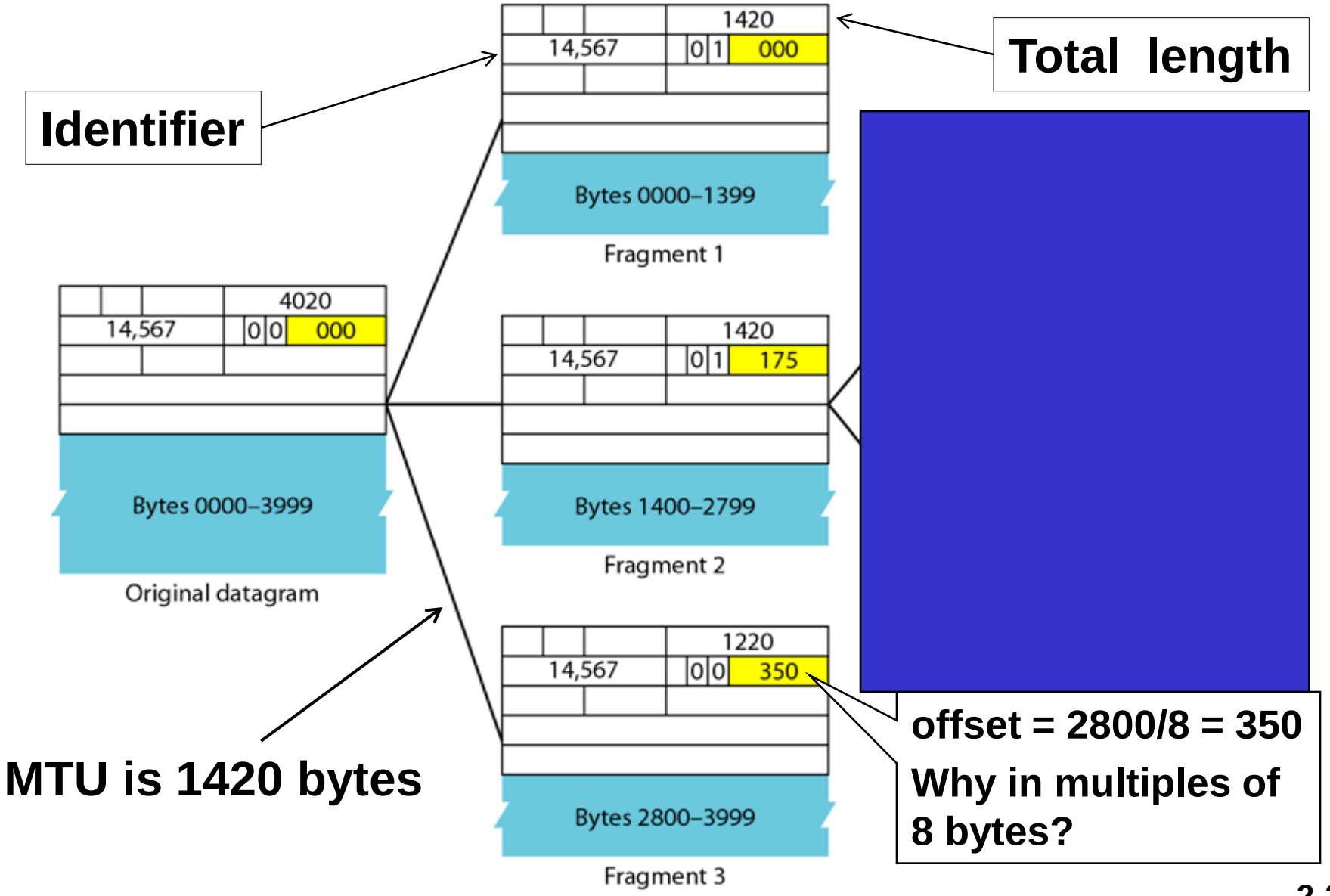
IPv4 Header – Fragmentation Fields



. . .

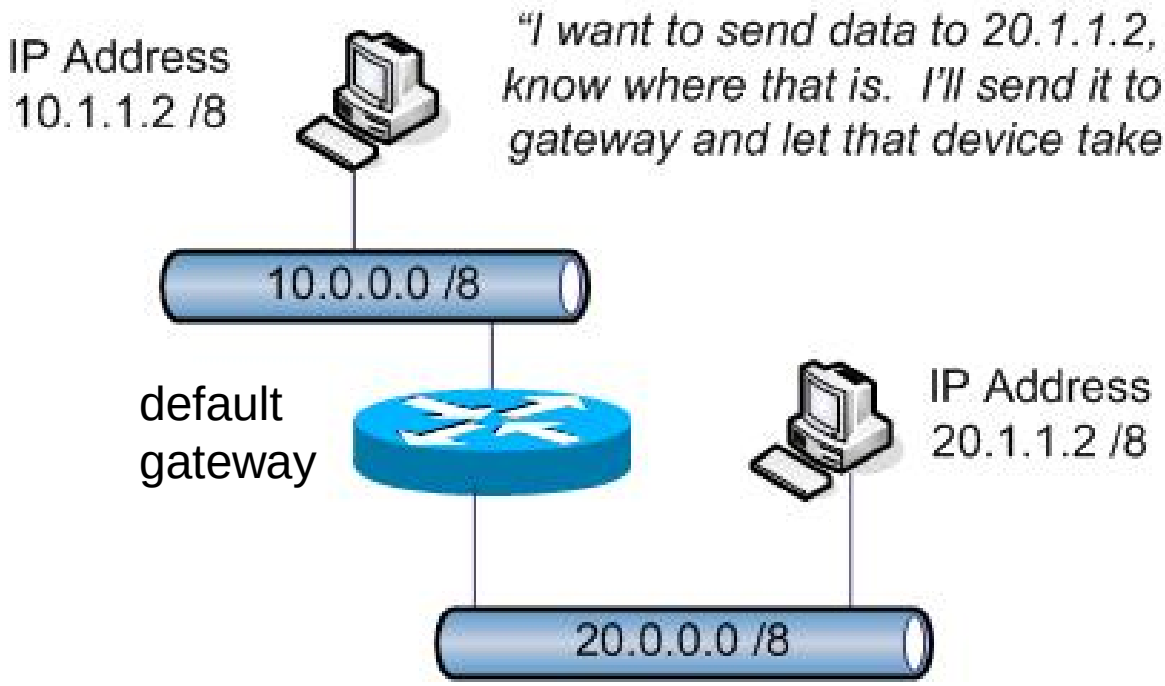
- Identification: For reassembly purpose, all fragments of a datagram contain the same identification value.
- Don't Fragment (DF): If the flag is set, the datagram is not fragmented.
- More Fragments (MF): The flag is set to 1 for all fragments except the last one.
- Fragment Offset: Tells where in the current datagram this fragment belongs. All fragments except the last one must be a multiple of 8 bytes, the basic fragment unit.

Example of IP Fragmentation



IP Routing

Now, we are ready to see how Internet Protocol works:



Typically, a host will not know how to send packets to destination outside its own network. Hence, it is configured with a default gateway (router) to assist in the forwarding.

Routing process

- **The routing table consist of the following**
  - **Network address: Destination Network address**
  - **Cost: Arbitrary cost, number of hops**
  - **Next hop : Who to past to next.**

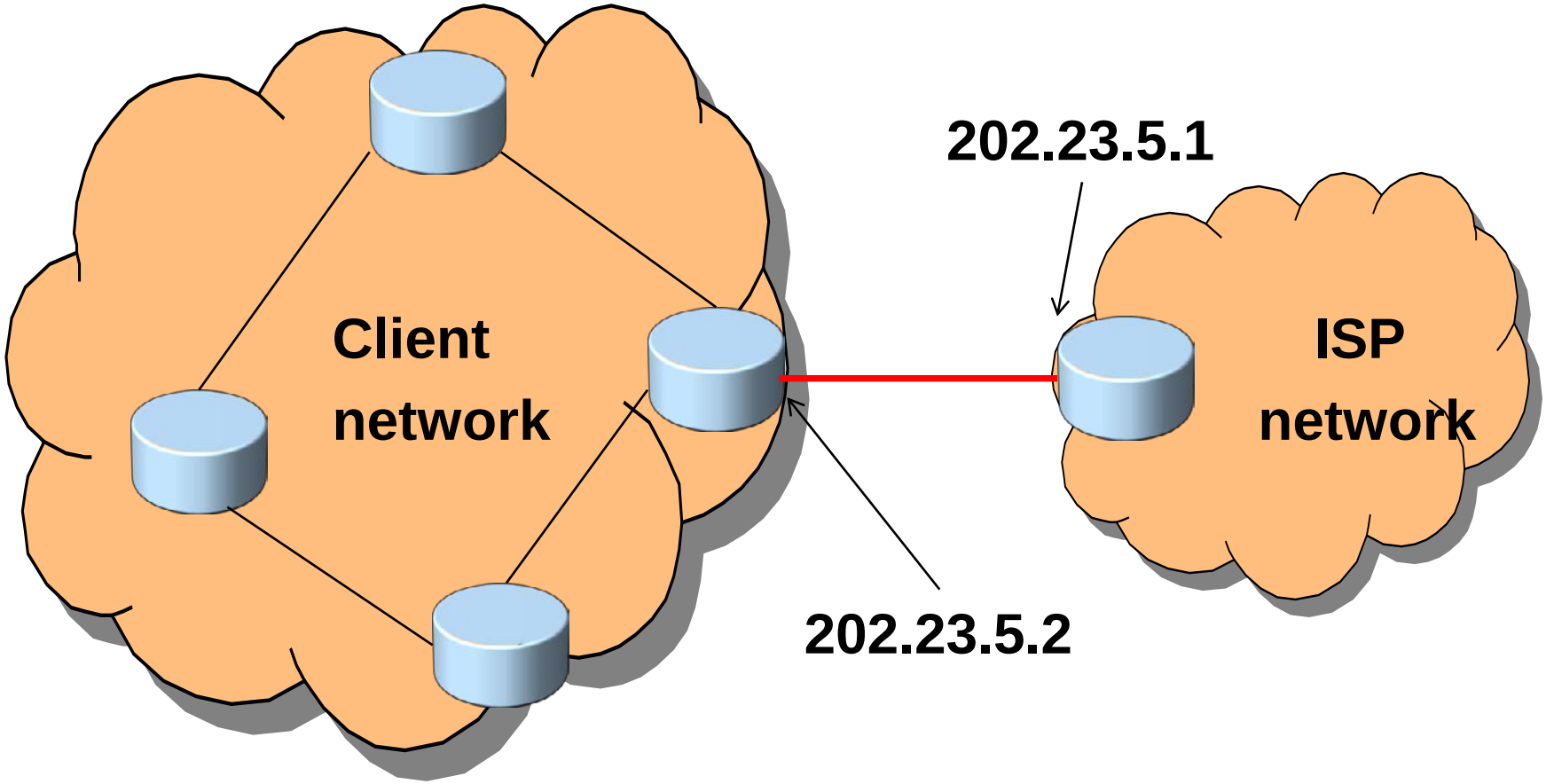
Network address	Cost	Next hop
155.69.0.0	1	Directly connected
122.0.0.0	1	Directly connected
194.8.9.0	7	122.5.6.1

Routing process

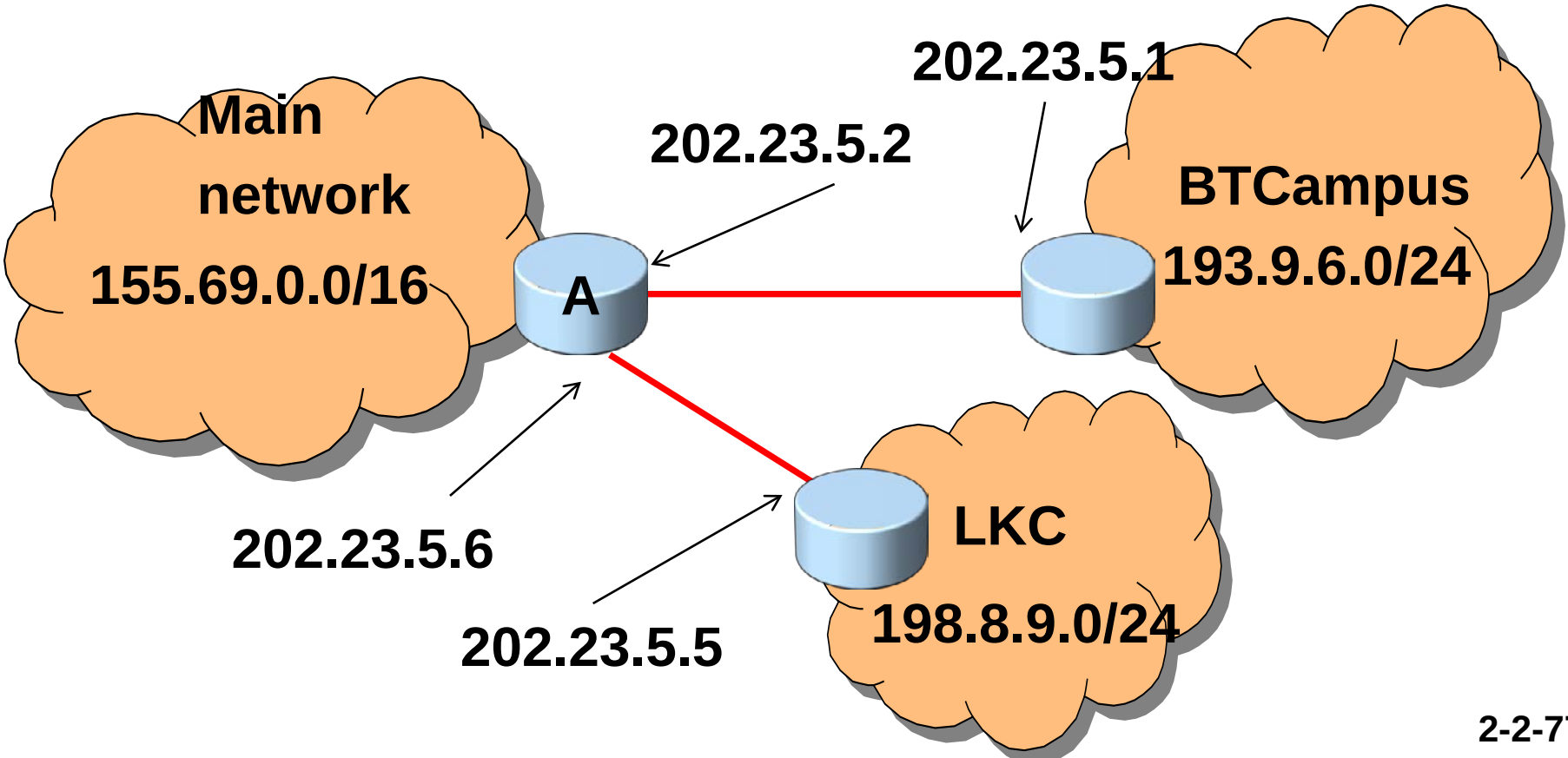
- **Extract Destination IP address and compute destination network address**
- **Loop-up next hop**
  - **IF**
    - **Destination IP network = Directly connected THEN send on the specified interface**
  - **ELSE IF**
    - **Destination IP address appear as host specific route THEN route as specified**
  - **ELSE IF**
    - **Destination Network address is in the routing table THEN send to the specified IP address through the specified interface**
  - **ELSE**
    - **Send packet through DEFAULT route**

Default route

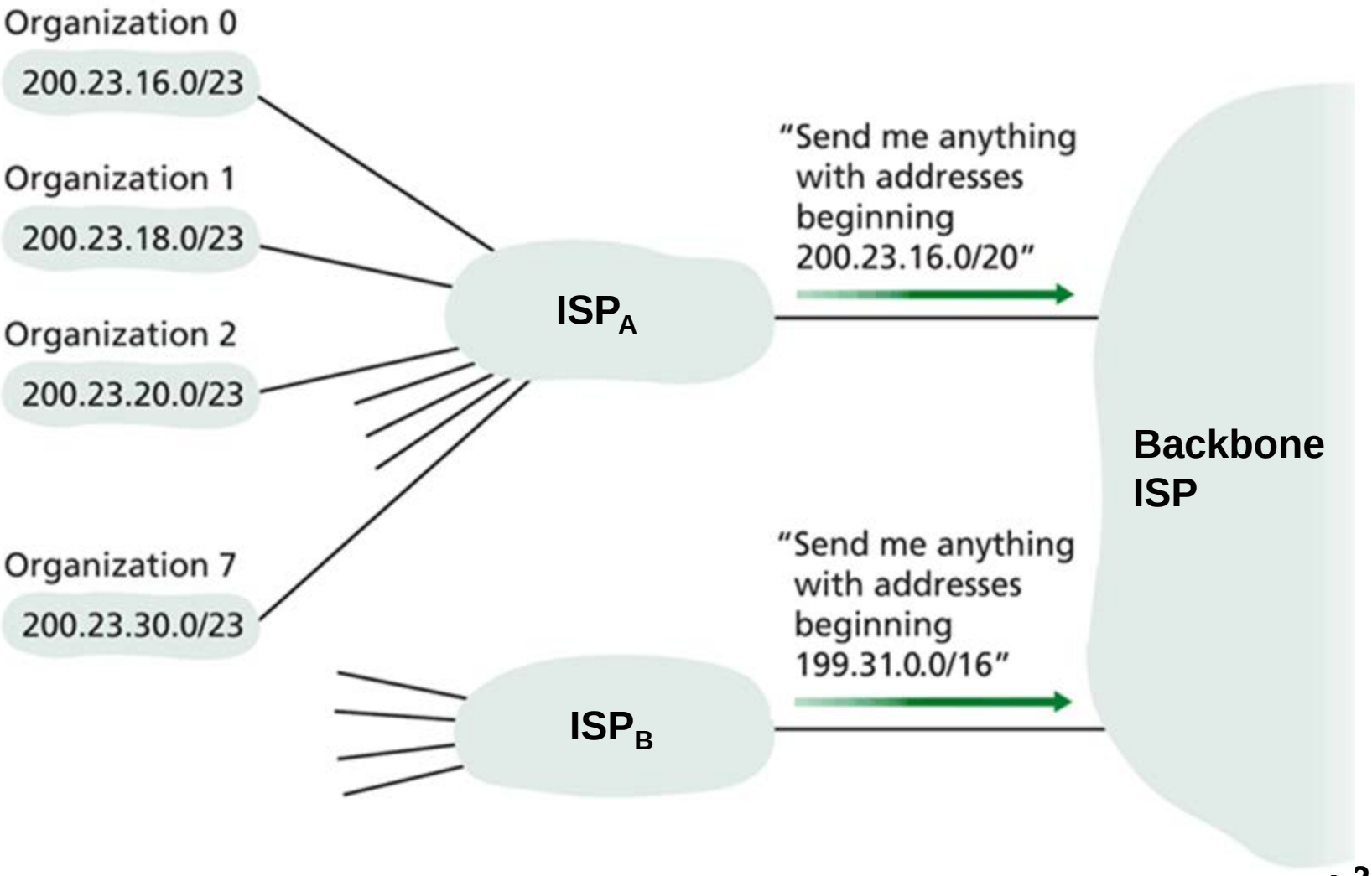
- **Default route will point to ISP router. 202.23.5.1**



Network address	Cost	Next hop
155.69.0.0	1	Directly connected
193.9.6.0	1	202.23.5.1
198.8.9.0	1	202.23.5.5



At routers, their routing tables are usually very large. Hence, the idea of address/route aggregation is introduced to reduce their sizes.



Address/route aggregation is achieved by combining multiple small prefixes into a single larger prefix called supernetting (in contrast with subnetting).

Organization 0:	200.23.00010000.0	200.23.16.0/23
Organization 1:	200.23.00010010.0	200.23.18.0/23
Organization 2:	200.23.00010100.0	200.23.20.0/23
Organization 3:	200.23.00010110.0	200.23.22.0/23
Organization 4:	200.23.00011000.0	200.23.24.0/23
Organization 5:	200.23.00011010.0	200.23.26.0/23
Organization 6:	200.23.00011100.0	200.23.28.0/23
Organization 7:	200.23.00011110.0	200.23.30.0/23

ISP Addr Block:

200.23.00010000.0

200.23.16.0/20

binary

Total Route Table Size 2015

Status Summary

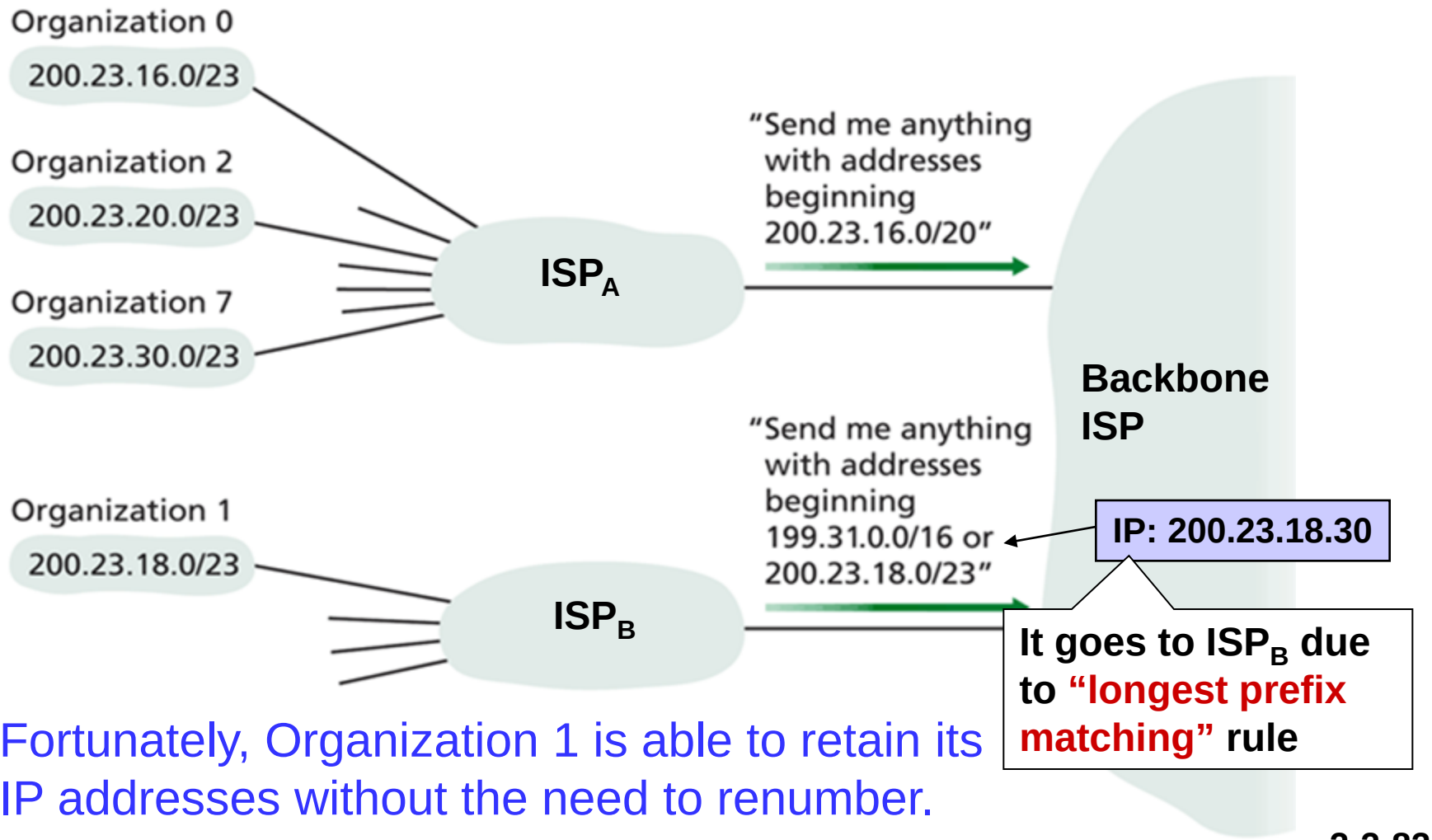
Table History

Date	Prefixes	CIDR Aggregated
22-02-15	539258	296450
23-02-15	539278	296680
24-02-15	539629	298059
25-02-15	540702	297505
26-02-15	538881	300039
27-02-15	541645	299462
28-02-15	541750	299479
01-03-15	541663	299467

Credit to Geoff Huston CIDR Report http://www.cidr-report.org/as2.0/

In addition, the idea of using longest prefix matching rule for routing is also being introduced.

What if Organization 1 switches to ISP_B?



Fortunately, Organization 1 is able to retain its IP addresses without the need to renumber.

How does longest prefix matching rule work?

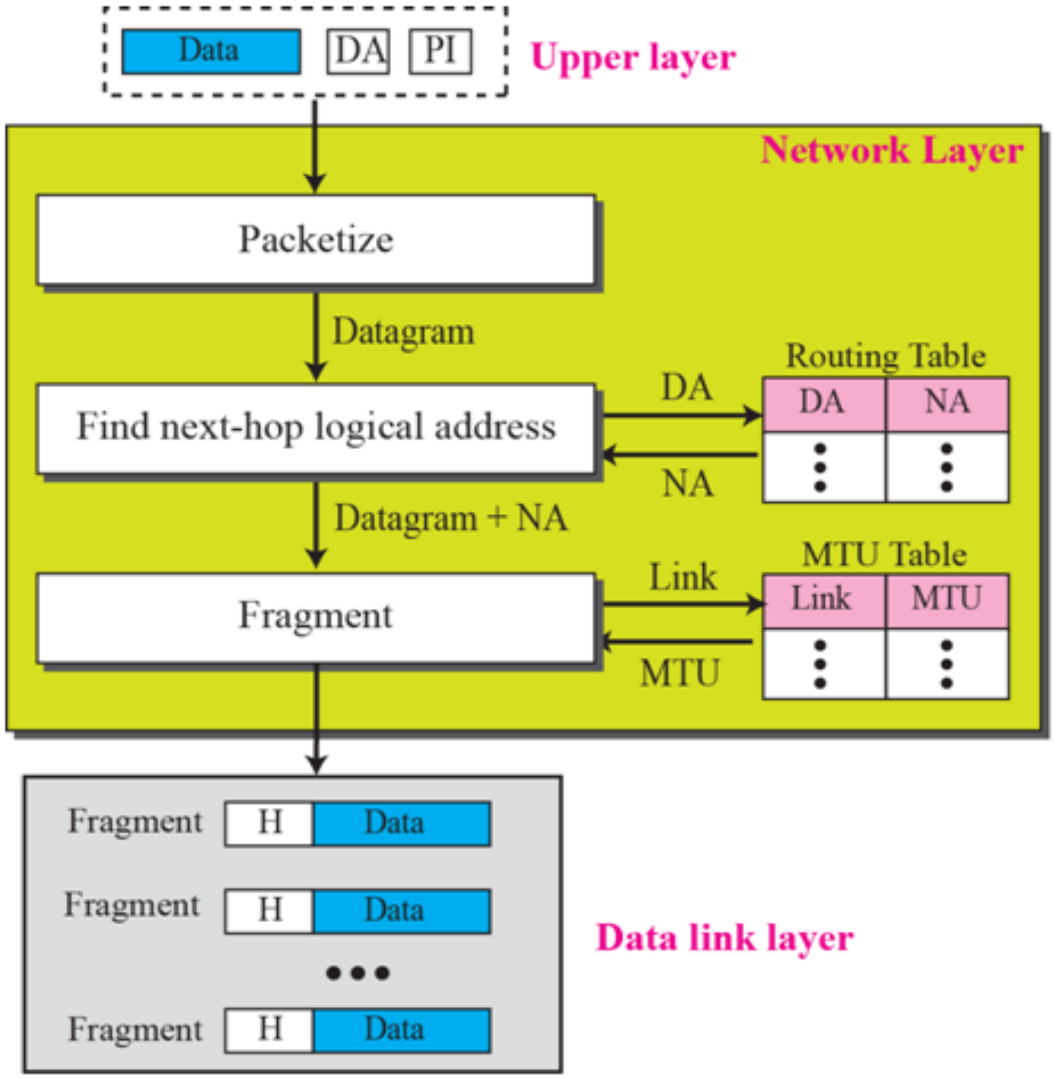
Backbone ISP's Routing Table:

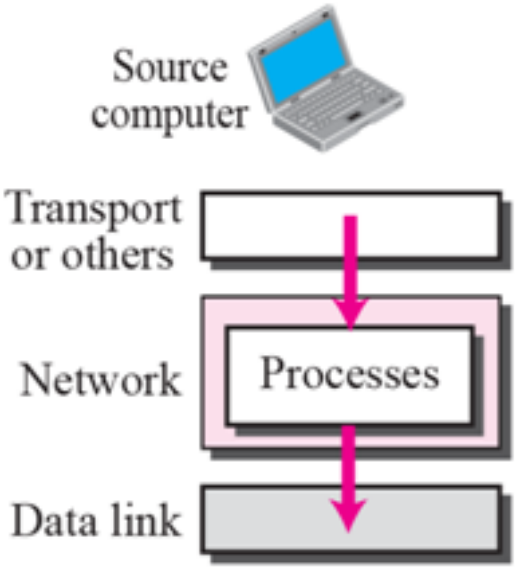
Network Destination	Netmask	Interface
200.23.00010000(16).0	255.255.11110000.0 (/20)	ISP_A
200.23.00010010(18).0	255.255.11111110.0 (/23)	ISP_B
. . .		

Match the longest prefix, so route to ISP_B

200.23.00010010(18).30 (Destination address)

Combining Routing and Fragmentation – At source host, IP encapsulates upper layer data into packet. Then determine route and MTU, and fragment if needed.

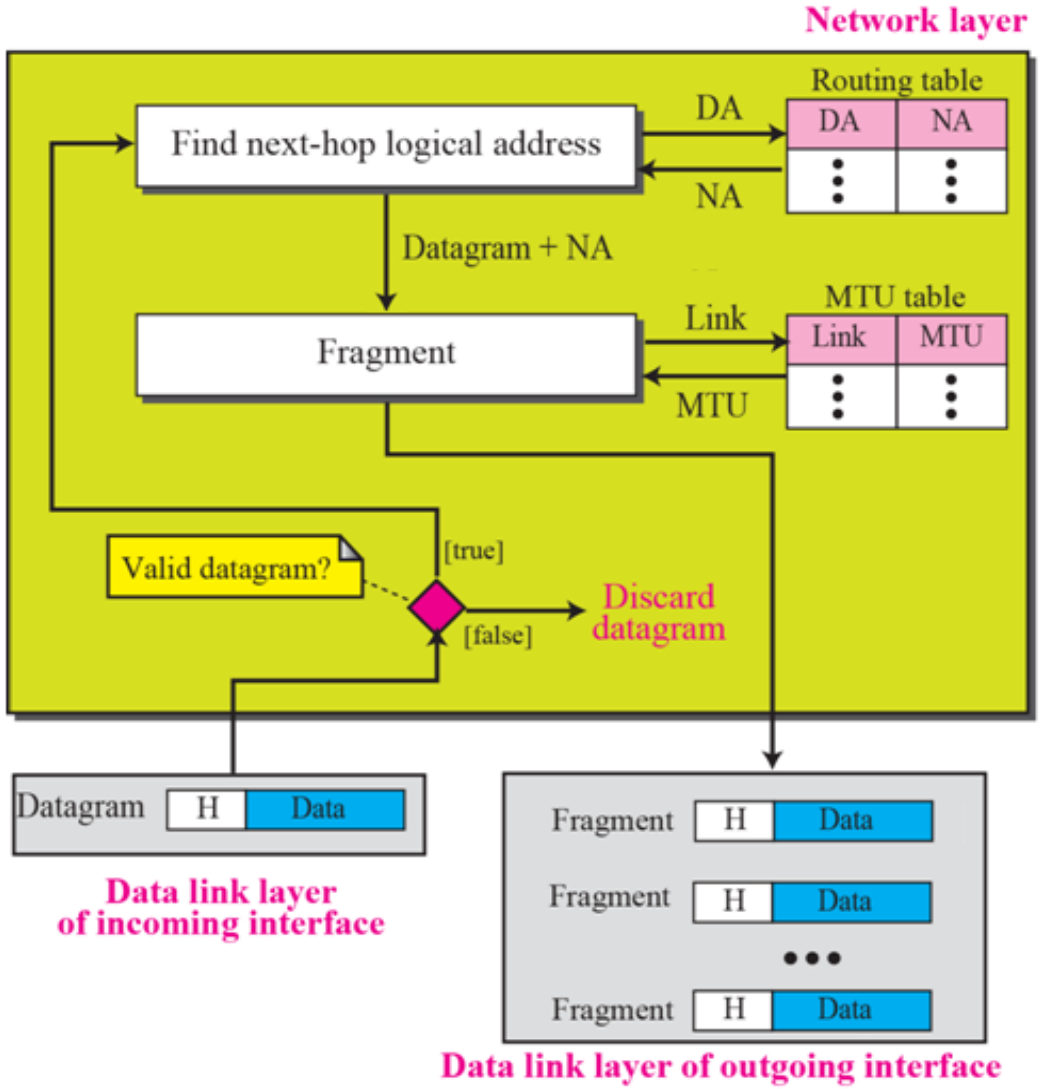


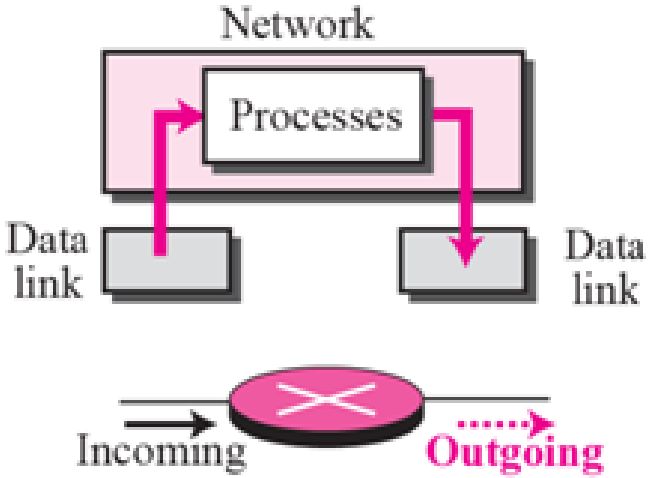


Legend

Data	Upper layer data
DA	Destination IP address
PI	Protocol ID
NA	Next-hop IP address
MTU	Maximum Transfer Unit
H	Datagram header

At each router, IP determines next route and MTU, and further fragment packets if necessary. But if DF flag is set, then it is simply discarded.

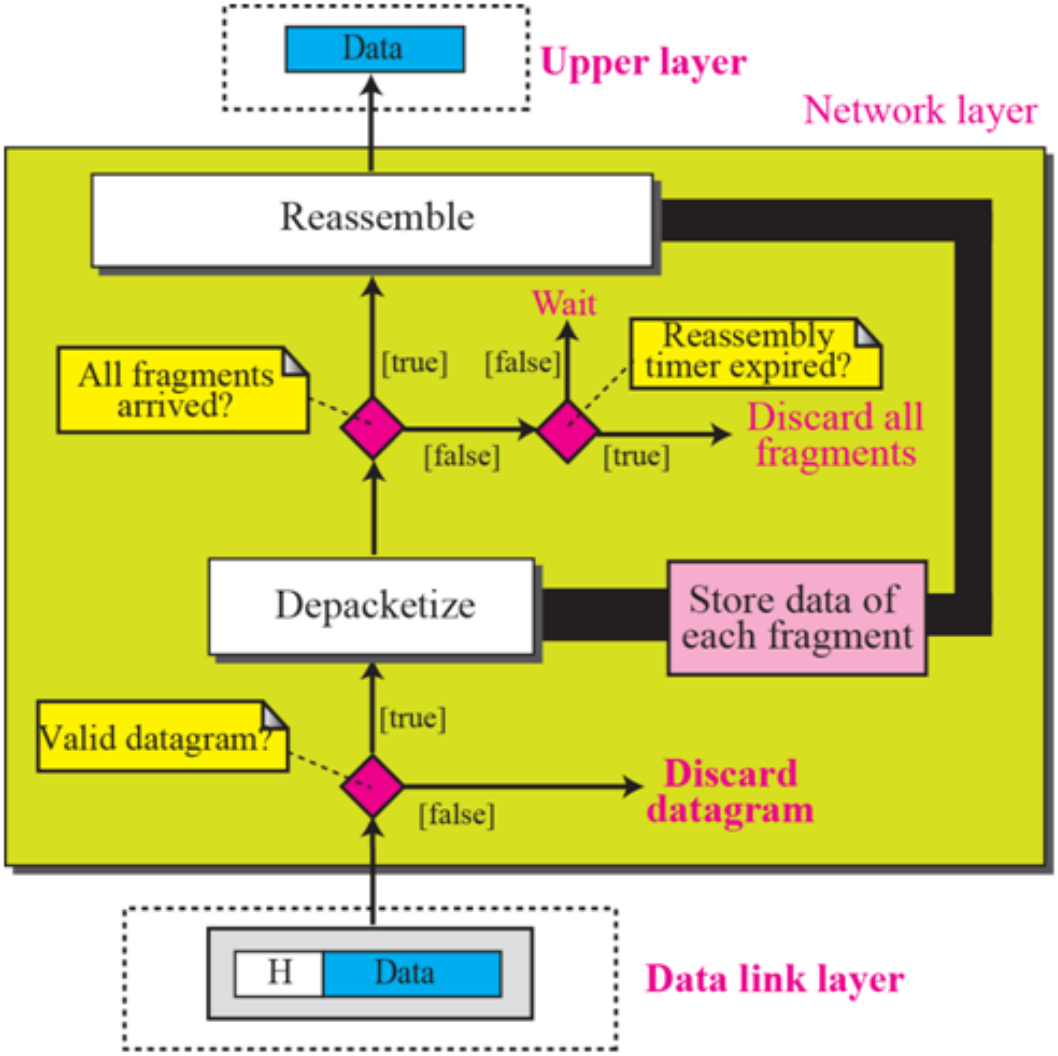


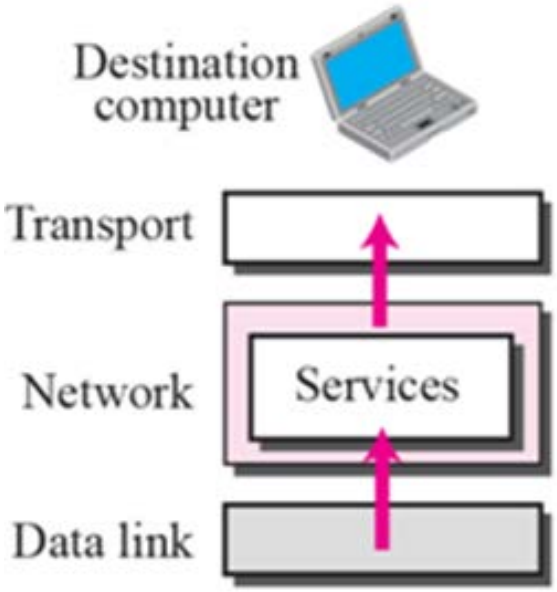


Legend

Data	Upper layer data
DA	Destination IP address
NA	Next-hop IP address
MTU	Maximum Transfer Unit
H	Datagram header

Finally, at destination host, IP reassembles packets (if fragmented) before returning data to upper layer.

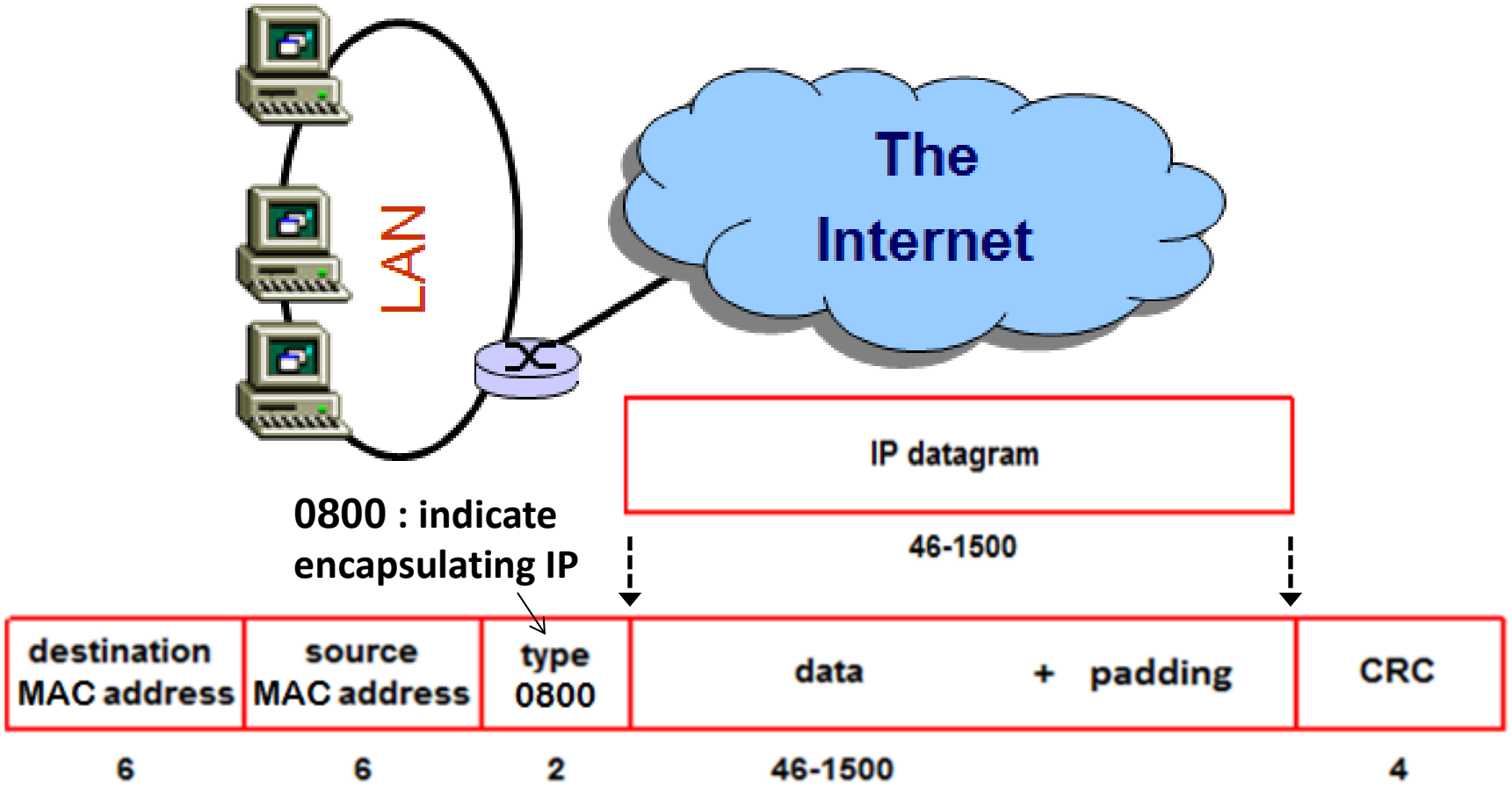




Legend

Data	Data of upper layer
H	Datagram header

To complete the picture, let's now consider IP over Ethernet (data link layer protocol).

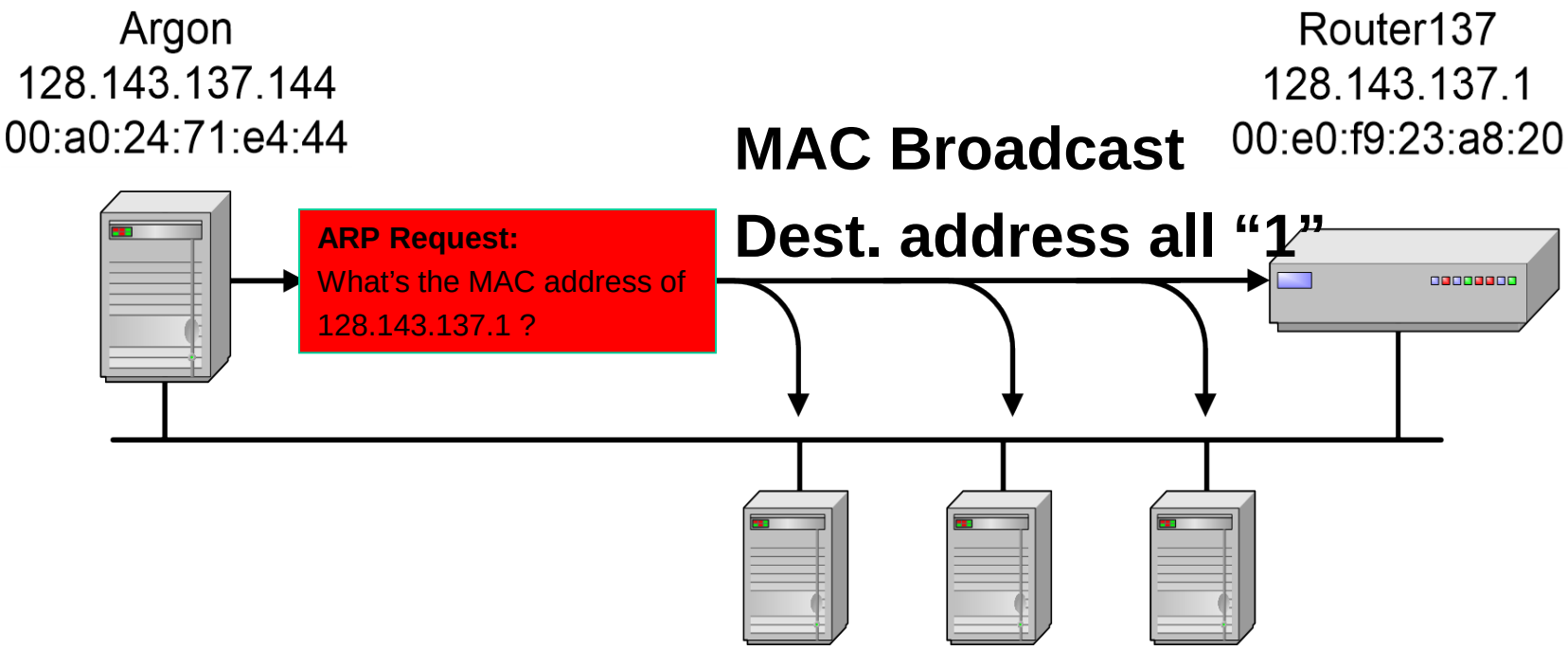


- **How to go from source to destination when only IP address is known but Ethernet requires MAC address?**

Address Resolution Protocol (RFC 826)

ARP Request:

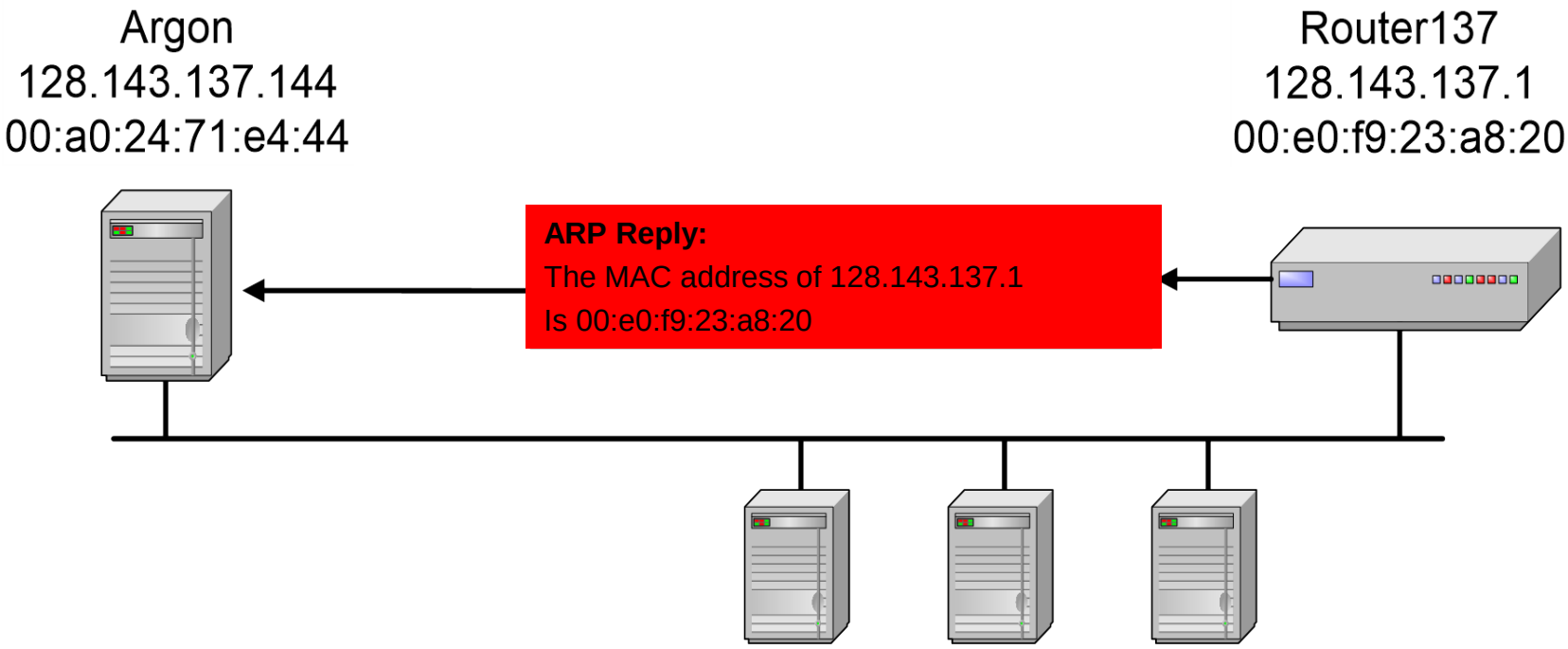
Argon broadcasts an ARP request to all stations on the network: "What is the MAC address of Router137?"



Address Resolution Protocol (RFC 826)

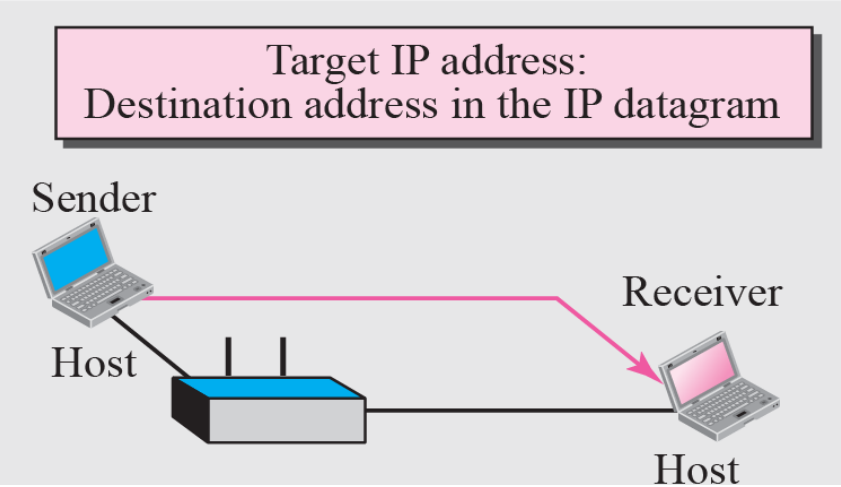
ARP Reply:

Router 137 responds with an ARP Reply which contains its MAC address

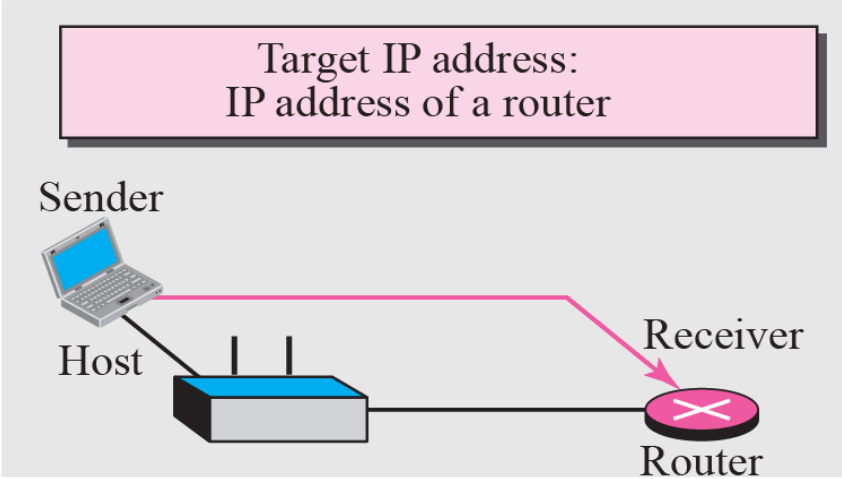


ARP Packet – Target IP Address

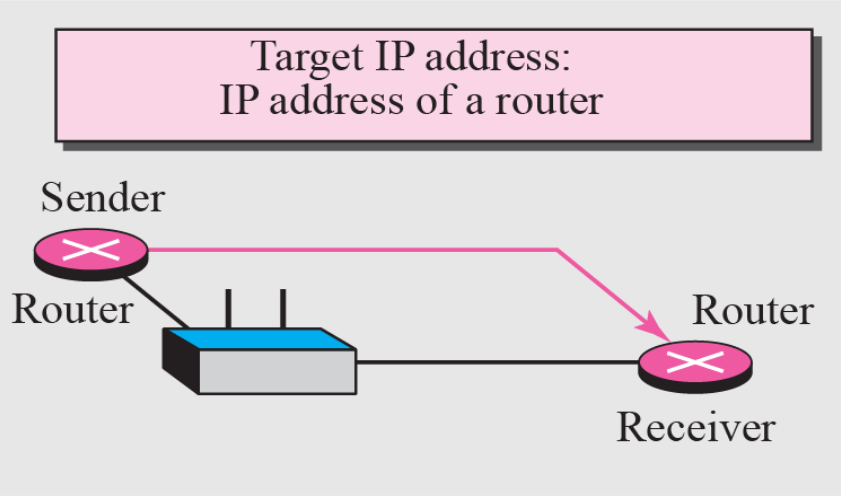
Case 1: A host has a packet to send to a host on the same network.



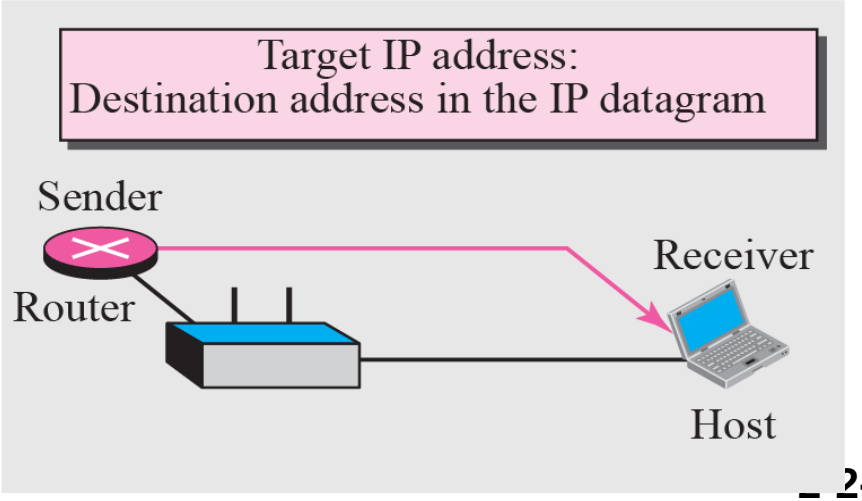
Case 2: A host has a packet to send to a host on another network.



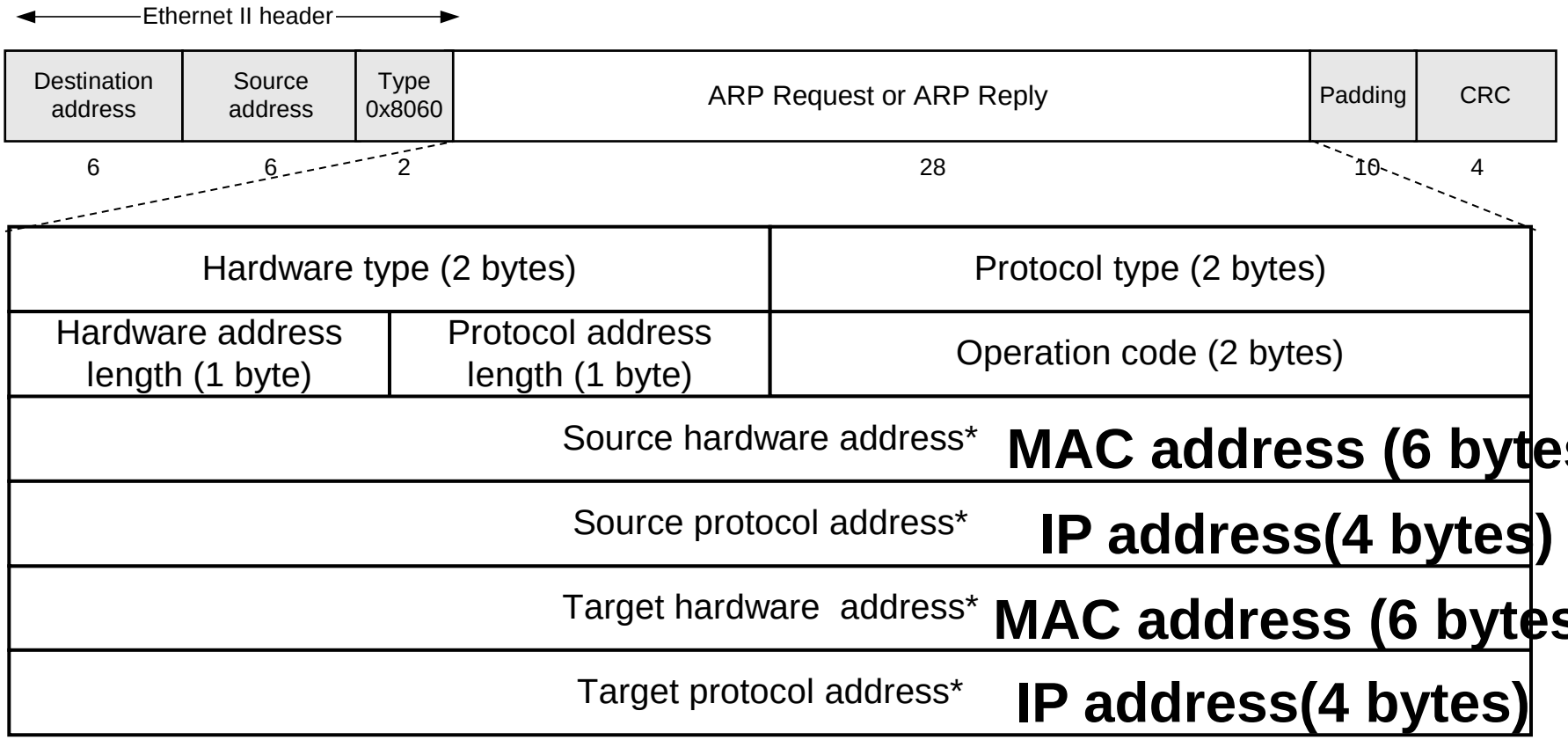
Case 3: A router has a packet to send to a host on another network.



Case 4: A router has a packet to send to a host on the same network.



ARP packet is sent directly over Ethernet frame:



* Note: The length of the address fields is determined by the corresponding address length fields

Hardware type: 0001_{16} for Ethernet **Protocol type**: 0800_{16} for IP

Operation code: 0001_{16} for request, 0010_{16} for reply

ARP Example

- *ARP Request from Argon(Client):*

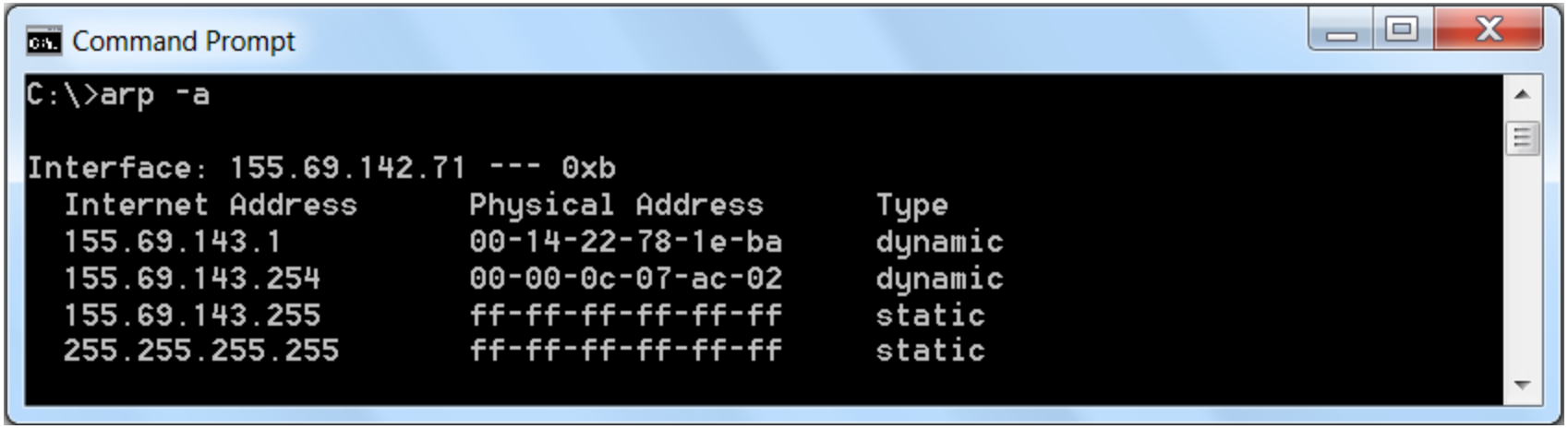
Source hardware address:	00:a0:24:71:e4:44
Source protocol address:	128.143.137.144
Target hardware address:	00:00:00:00:00:00
Target protocol address:	128.143.137.1

- *ARP Reply from Router137:*

Source hardware address:	00:e0:f9:23:a8:20
Source protocol address:	128.143.137.1
Target hardware address:	00:a0:24:71:e4:44
Target protocol address:	128.143.137.144

ARP Cache

- Since sending an ARP request/reply for each IP datagram is inefficient, hosts maintain a cache (ARP Cache) of current entries. Typically, the entries are configured to expire after 2-20 minutes.



155.69.143.254 is the gateway/router address

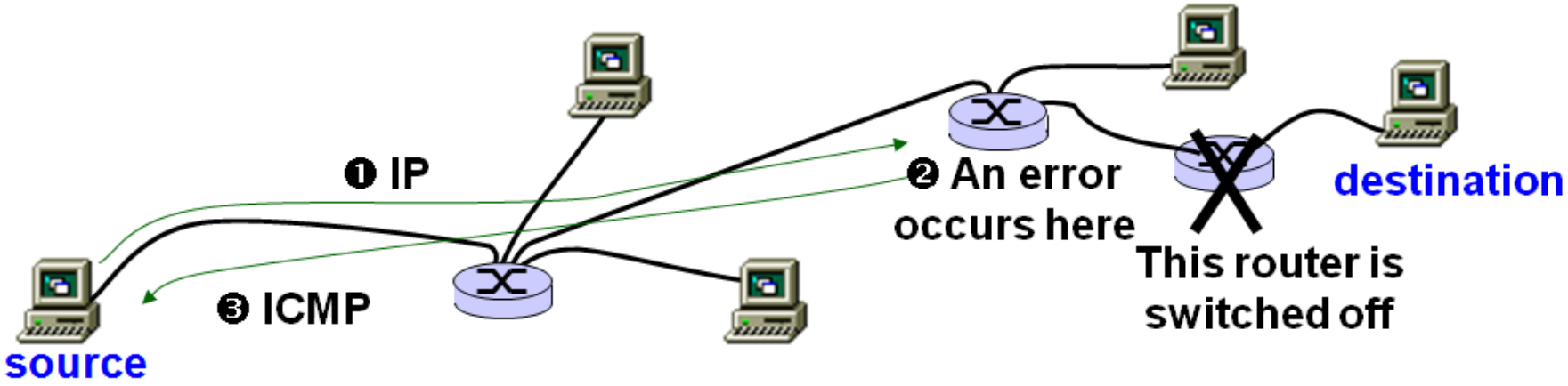
IP Related Protocols

ICMP, PING, TRACERT

Internet Control Message Protocol (RFC 792 & 1122)

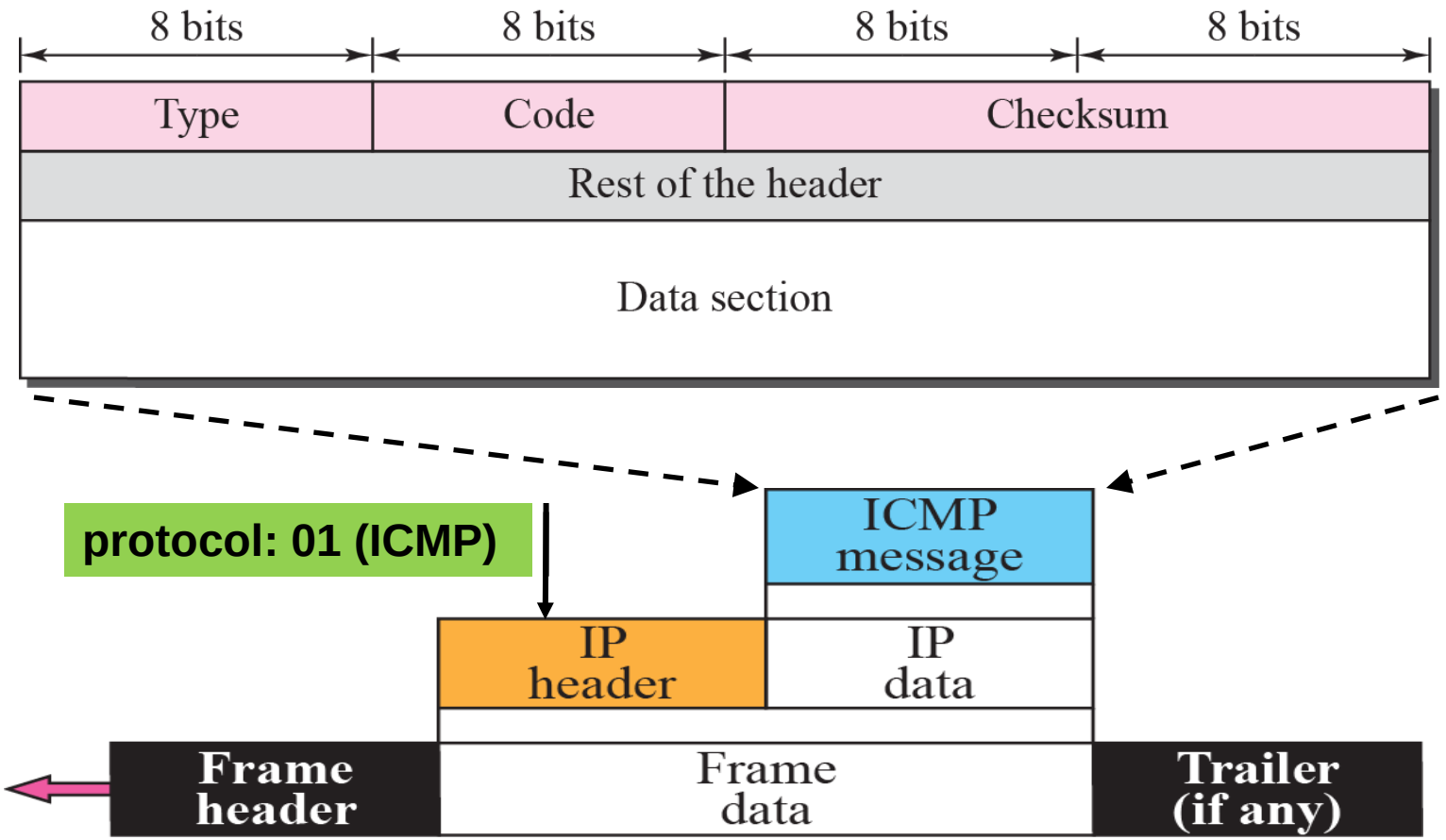
What if a source IP packet cannot be delivered to destination?

Although IP does not perform error control, it must still report errors. Otherwise, the source TCP (transport layer) may retransmit all data causing IP to send same packets again that can never reach the destination.



Hence, ICMP is created for routers/hosts to report errors to the source. It is the responsibility of the source to handle the reported problems.

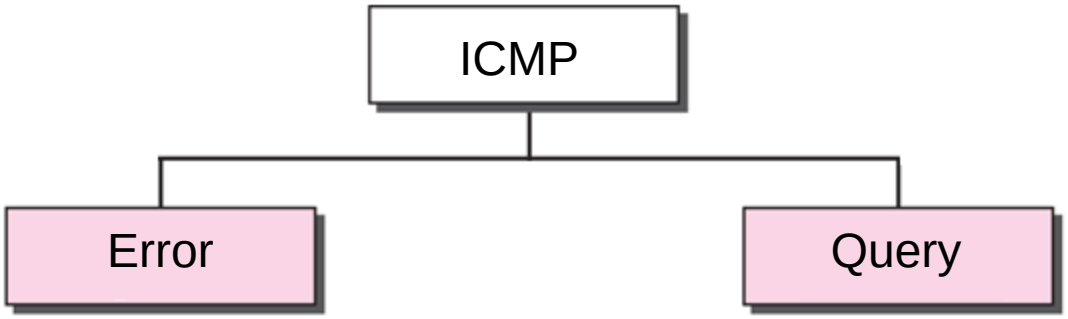
ICMP packet is sent over IP packet, which in turn is sent over data link layer protocol, e.g. Ethernet.



Note: Although ICMP is over IP, it is considered a layer-3 protocol.

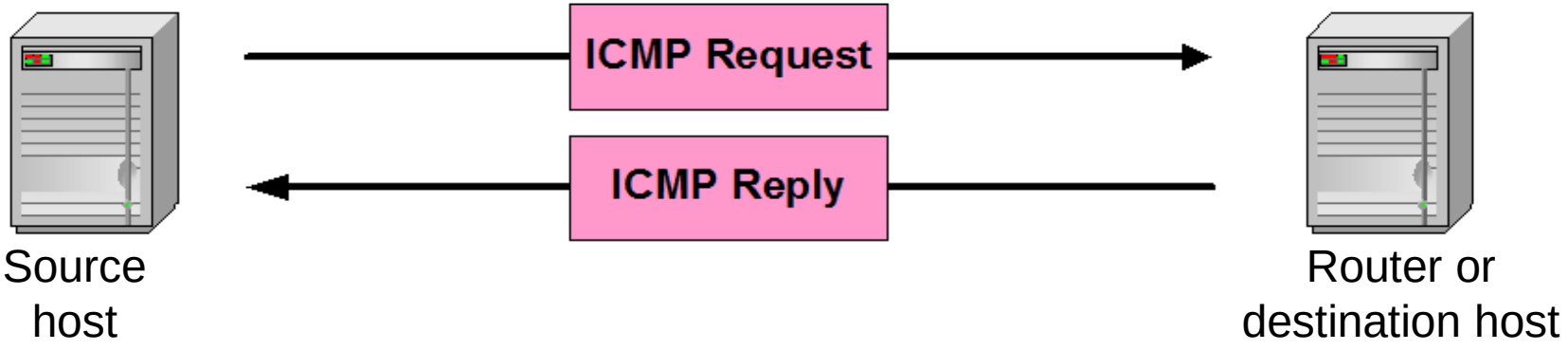
ICMP

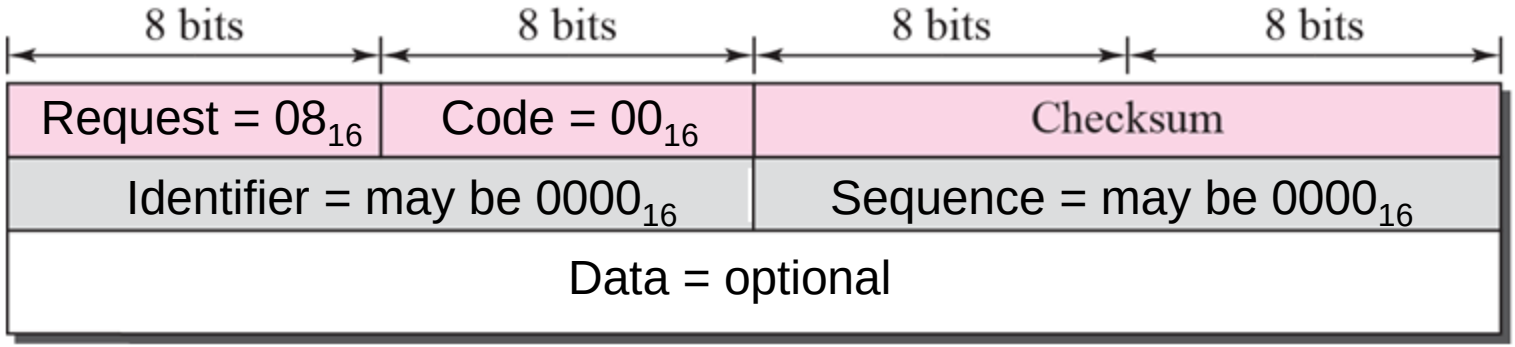
Generally, there are 2 types of ICMP messages:

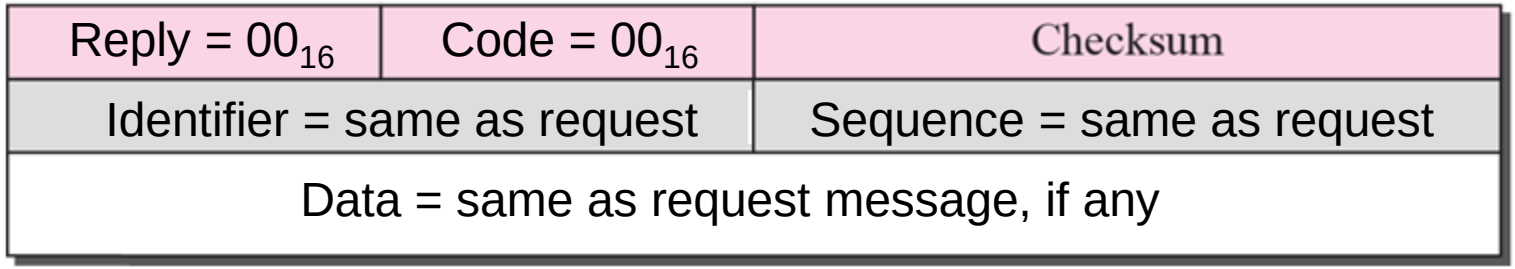


Category	*Type*	*Message*
Error-reporting messages	3	Destination unreachable
	4	Source quench
	11	Time exceeded
	12	Parameter problem
	5	Redirection
Query messages	8 or 0	Echo request or reply
	13 or 14	Timestamp request or reply

ICMP echo request message is sent by a source host to query a router or destination host, which will respond with an ICMP echo reply message.

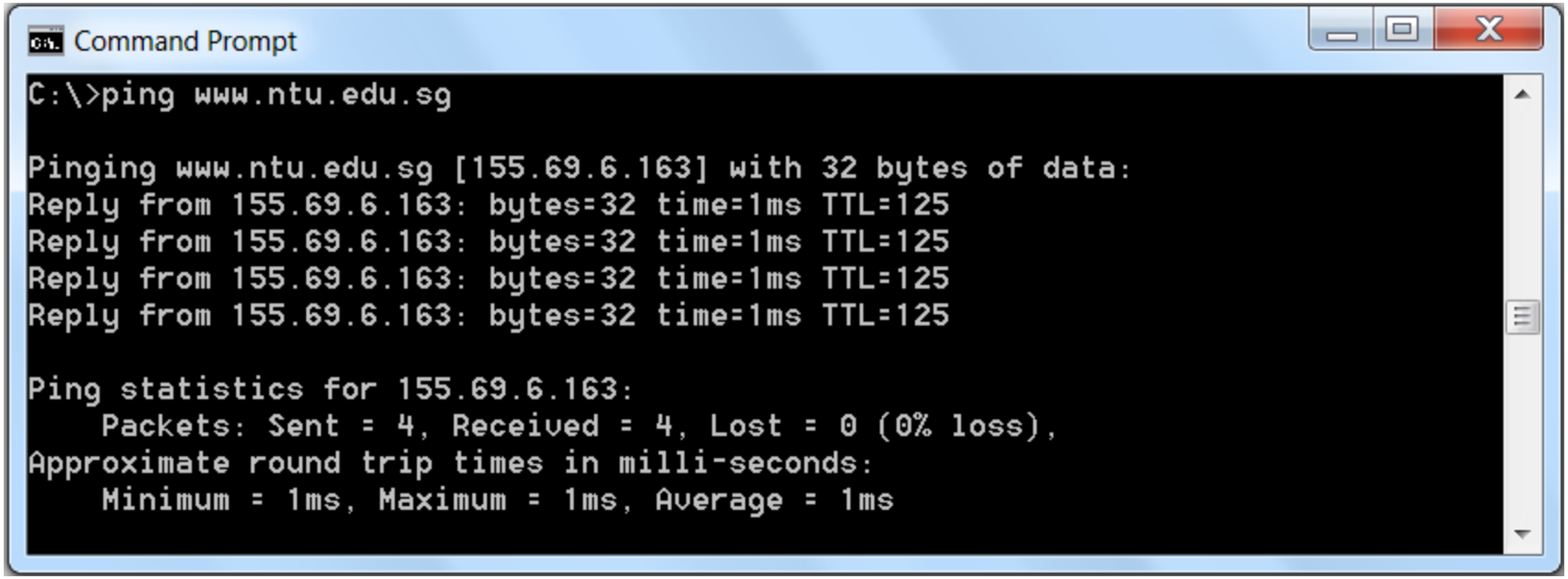




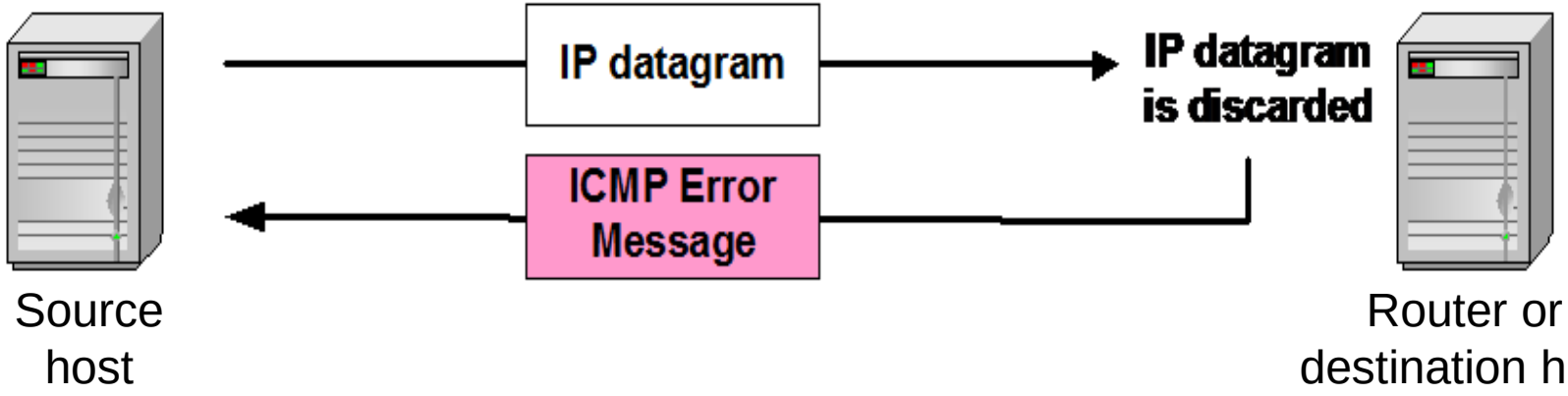


ping network tool

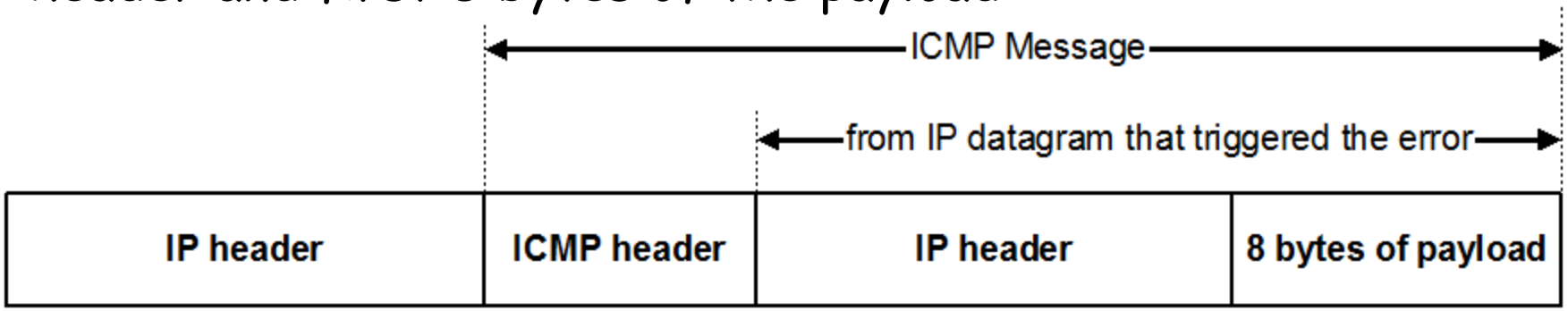
ping (Packet InterNet Grouper) is a useful network debugging tool for testing the reachability of a host/router. Basically, it operates by sending/receiving ICMP echo messages.



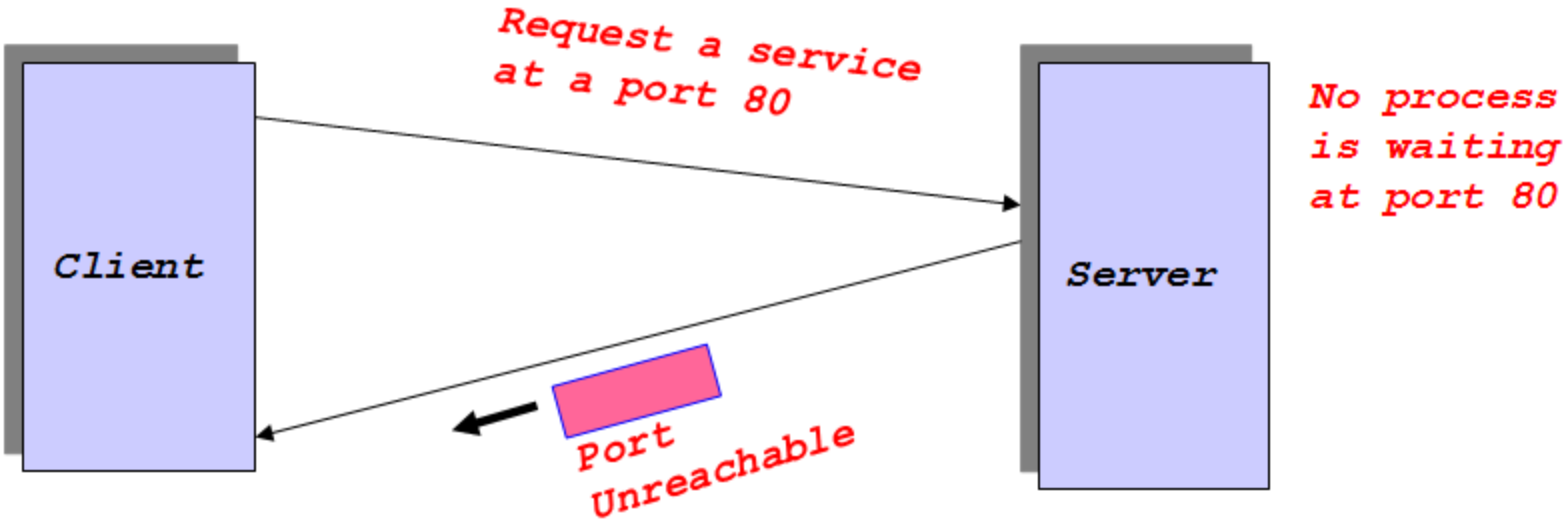
ICMP error message is sent by a router or destination host to inform source host that its datagram has been received in error and discarded.

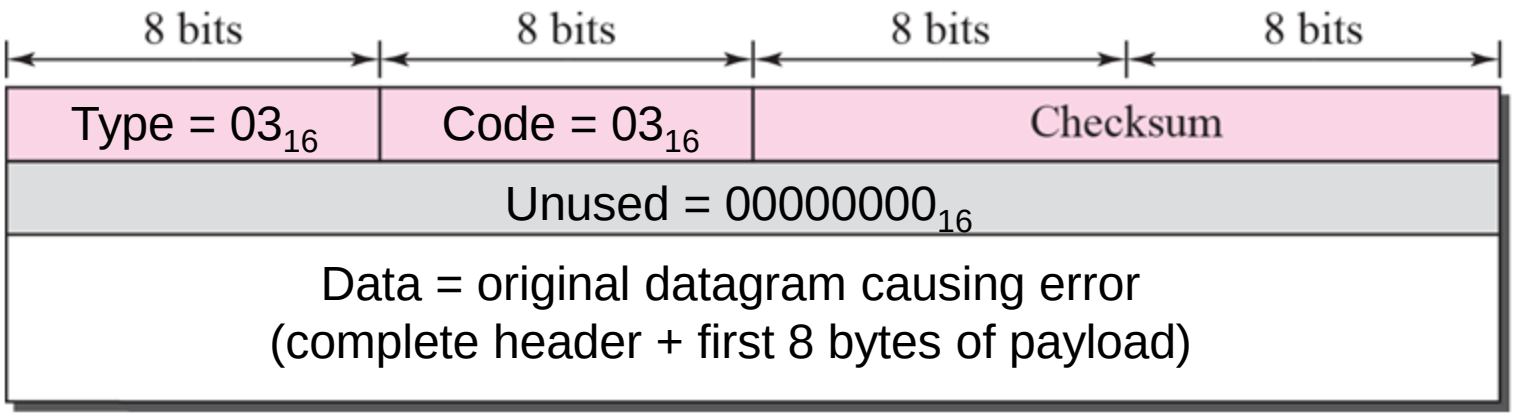


The data section of ICMP error message will contain part of the original IP datagram in error - the complete IP header and first 8 bytes of the payload:



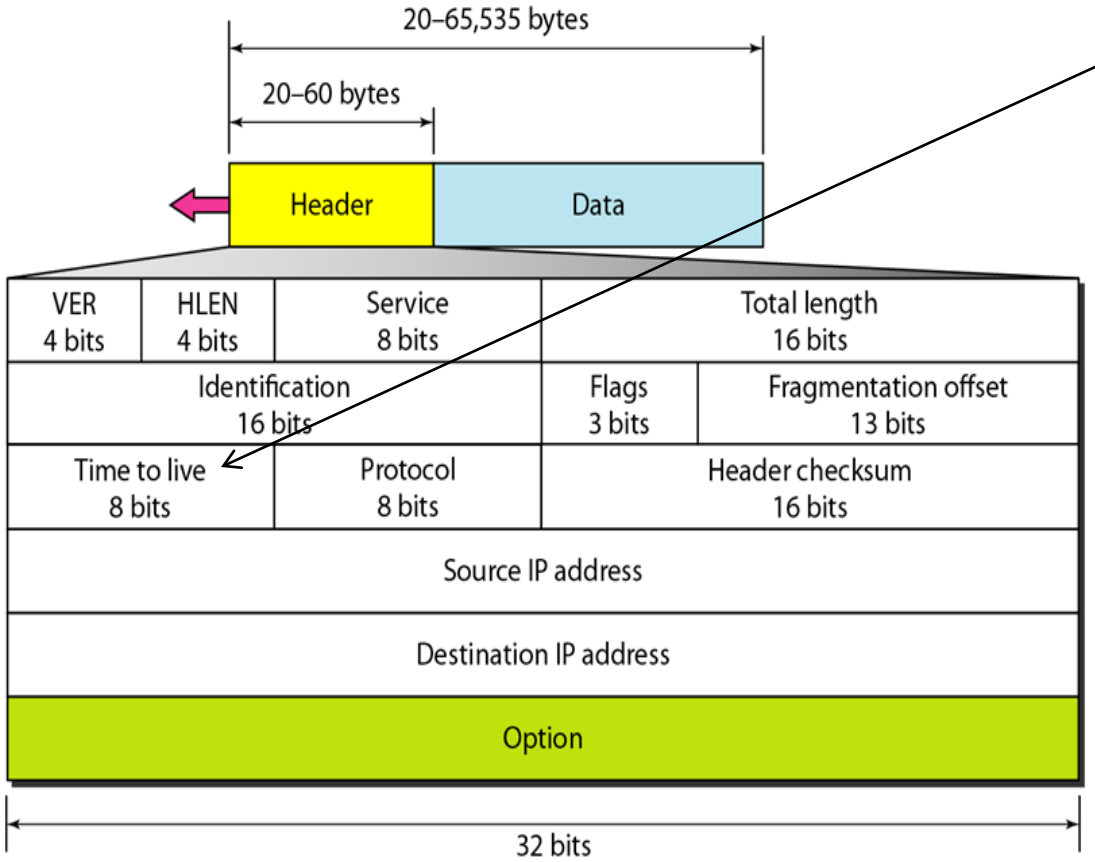
Example: ICMP Port Unreachable





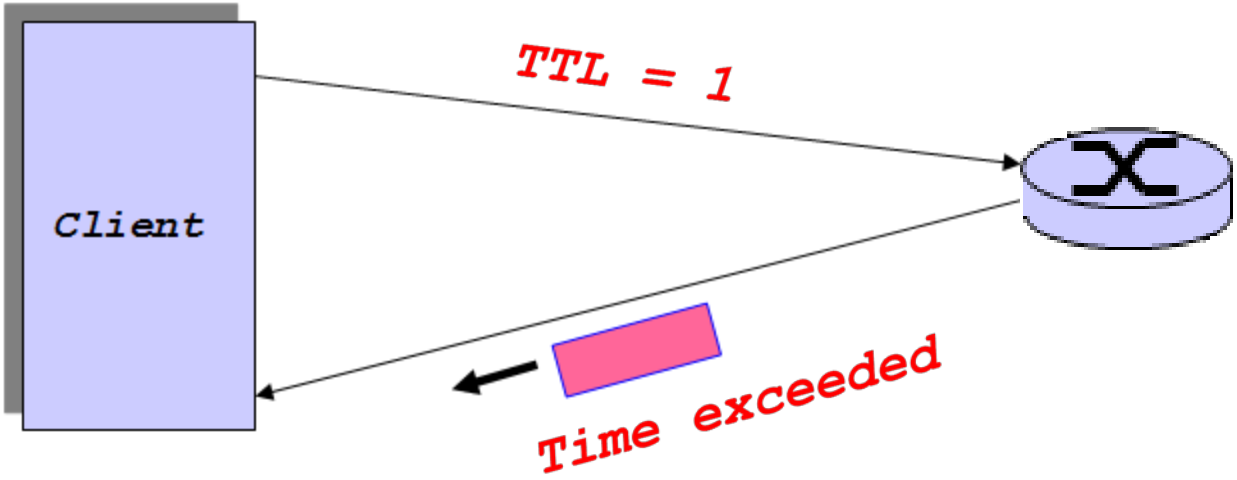
Example: ICMP Time Exceeded

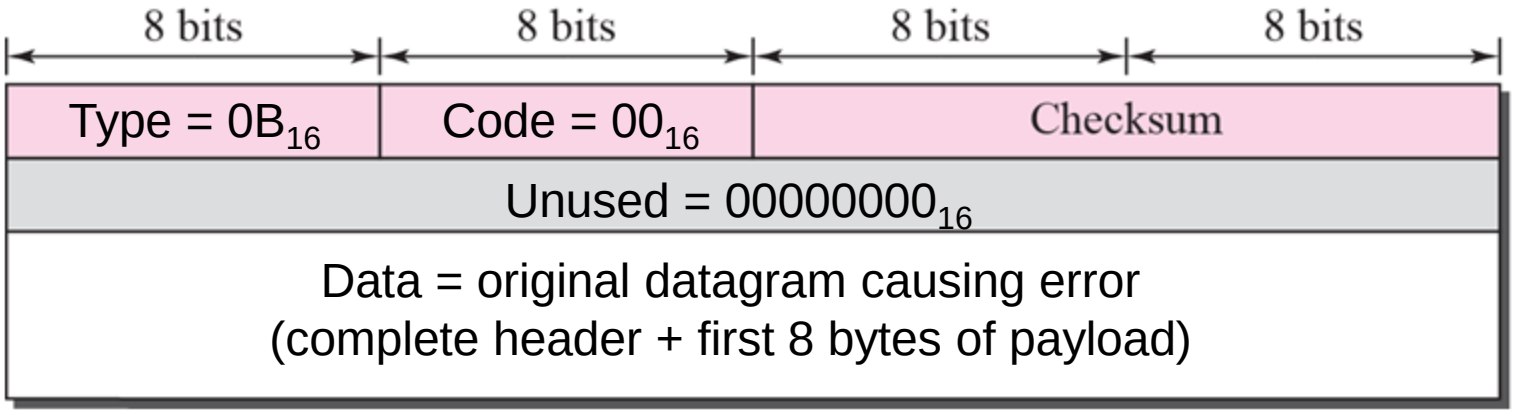
Finally, last field in IP header that we've yet to discussed :)



- Time to Live: It is a counter used to limit the IP datagram lifetime, number of hops. The counter is initialized with an integer value up to 255, and when it reaches zero, the IP datagram is discarded.

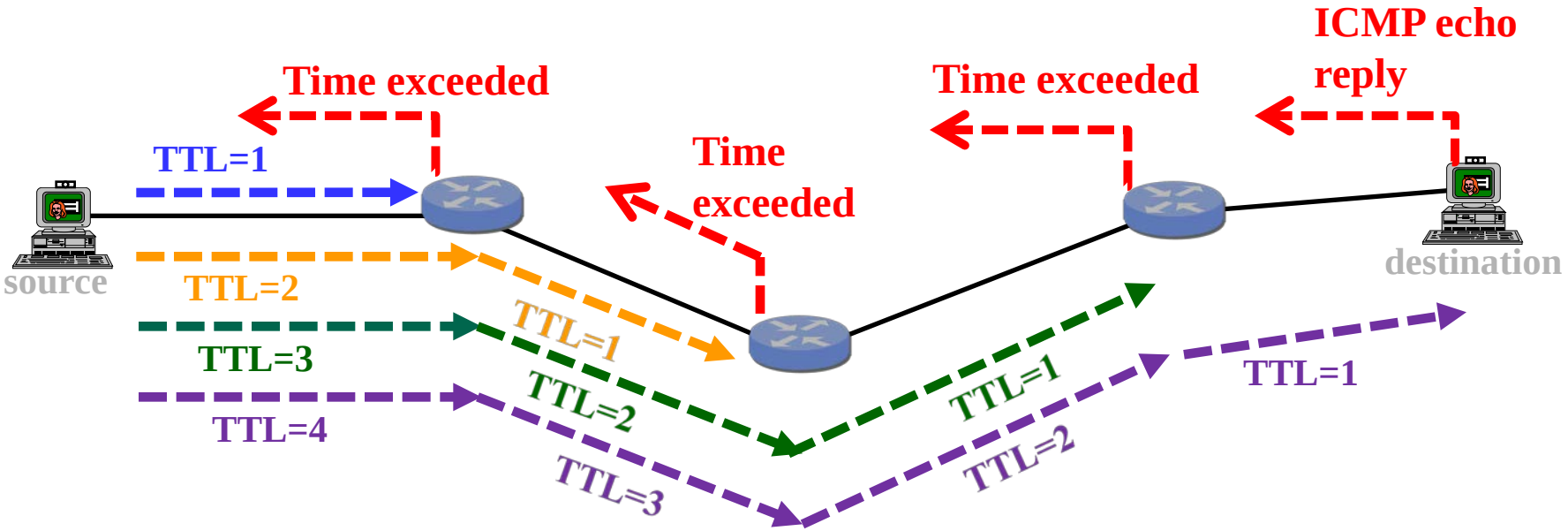
Example: ICMP Time Exceeded





tracert network tool

tracert (trace route) is another useful network debugging tool for tracing a path from source to destination host.



It operates by sending a sequence of ICMP echo request over IP with TTL set to 1, 2, … until the destination is reached.

Note: tracert can also be implemented using UDP (layer-4) with unused port number (e.g. 33534) over IP.

tracert network tool

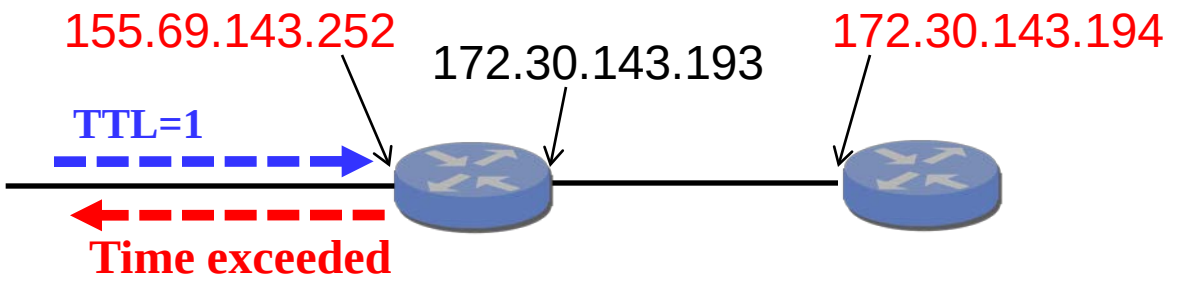
```
Command Prompt
C:\>tracert www.ntu.edu.sg

Tracing route to www.ntu.edu.sg [155.69.6.163]
over a maximum of 30 hops:

  1    <1 ms    <1 ms    <1 ms  155.69.143.252
  2     1 ms     1 ms     1 ms  172.30.143.194
  3     1 ms     1 ms     1 ms  172.30.2.193
  4     1 ms     1 ms     1 ms  www.ntu.edu.sg [155.69.6.163]

Trace complete.
```

Recall IP address is associated with an interface, not a device. Typically, the source address field of the IP that carries the ICMP message is the address of the interface that sends it.



Singapore to Switzerland

traceroute to www.isg.hest.ethz.ch (129.132.19.217), 30 hops max, 60 byte packets

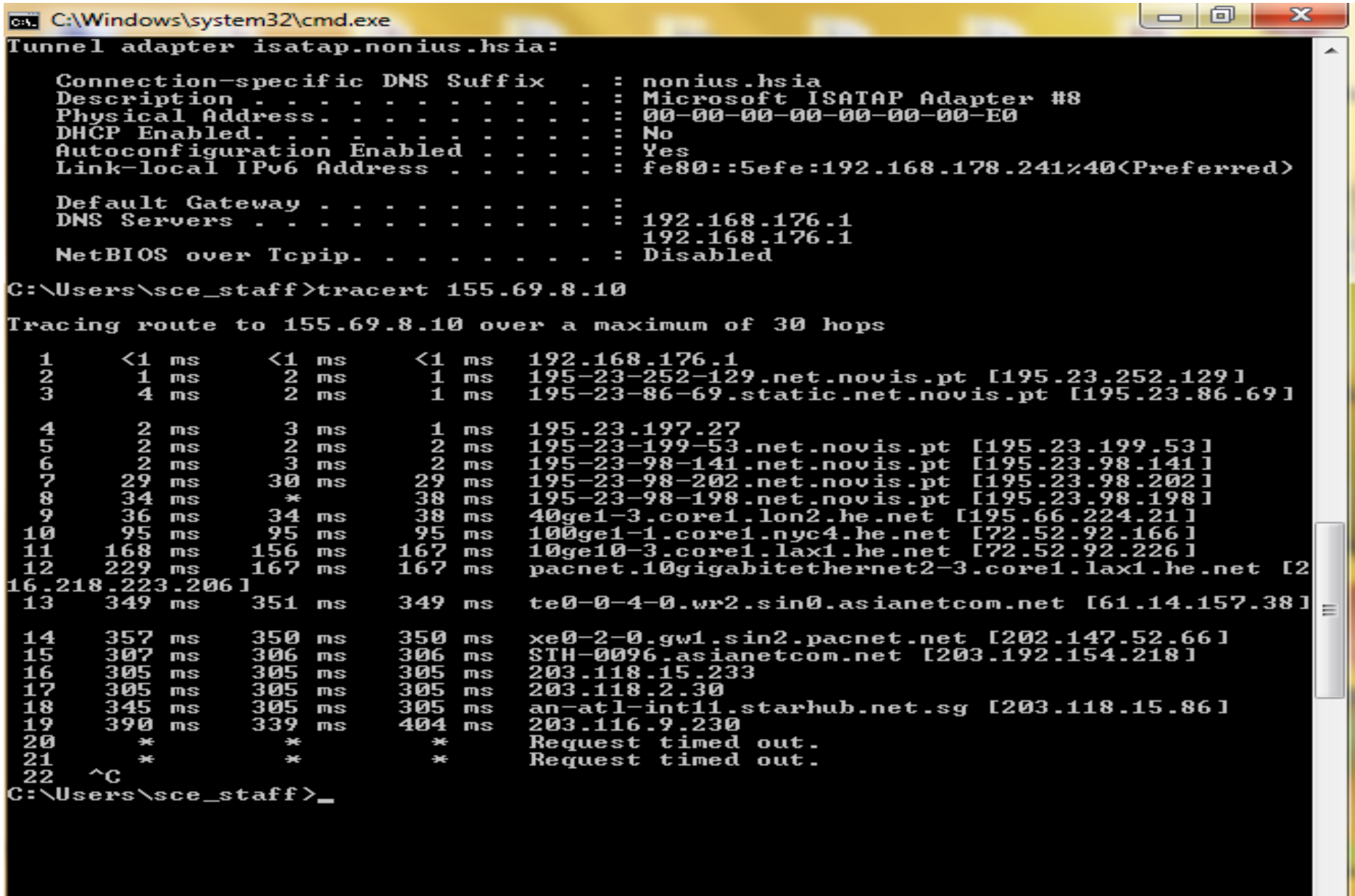
1 203.30.39.254 (203.30.39.254) 0.843 ms 1.023 ms 0.804 ms
2 sg-ge-01-v4.bb.tein3.net (202.179.249.57) 0.958 ms 0.957 ms 0.859 ms
3 mb-so-01-v4.bb.tein3.net (202.179.249.82) 58.638 ms 58.545 ms 58.563 ms
4 eu-mad-pr-v4.bb.tein3.net (202.179.249.86) 171.371 ms 171.378 ms 171.353 ms
5 as2.rt1.gen.ch.geant2.net (62.40.112.25) 193.451 ms 193.471 ms 193.444 ms
6 switch-lb2-gw.rt1.gen.ch.geant.net (62.40.124.106) 193.580 ms 193.627 ms 193.425 ms
7 swiLS2-10GE-1-3.switch.ch (130.59.37.2) 194.444 ms 194.408 ms 194.485 ms
8 swiEZ2-10GE-1-1.switch.ch (130.59.36.206) 197.782 ms 197.859 ms 197.806 ms
9 rou-gw-rz-tengig-to-switch.ethz.ch (192.33.92.1) 197.818 ms 197.834 ms 197.811 ms
10 rou-fw-rz-rz-gw.ethz.ch (192.33.92.169) 197.883 ms 198.359 ms 200.217 ms
11 * * *

Singapore to Brazil

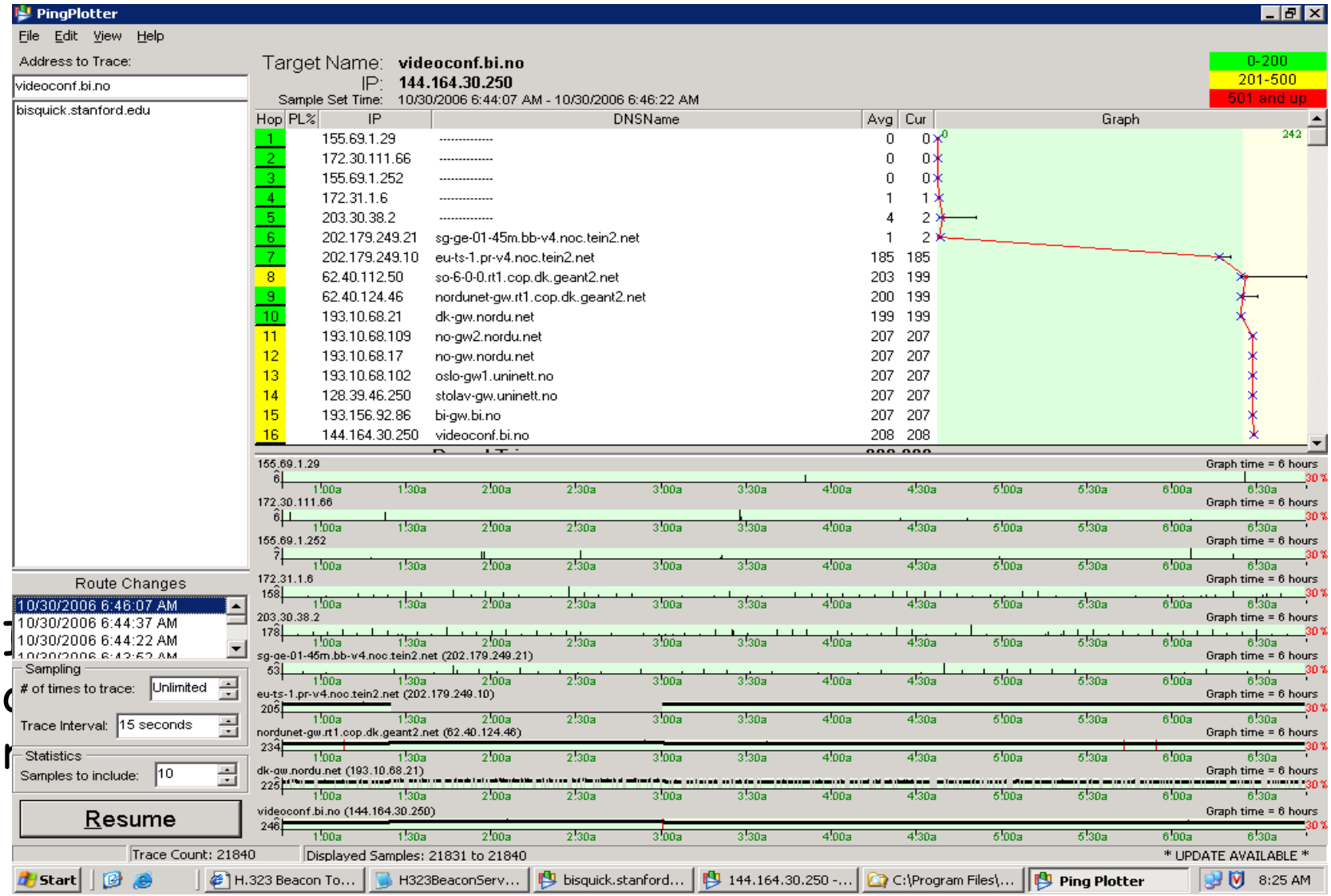
traceroute to www.rnp.br (200.143.193.5), 64 hops max

- 1 203.30.39.254 (203.30.39.254) 0.457ms 0.340ms 0.409ms
- 2 192.31.99.85 (192.31.99.85) 0.742ms 0.345ms 0.430ms
- 3 192.31.99.249 (192.31.99.249) 179.086ms 179.436ms 180.342ms
- 4 192.31.99.161 (192.31.99.161) 229.212ms 232.749ms 234.068ms
- 5 192.31.99.134 (192.31.99.134) 237.848ms 223.930ms 223.785ms
- 6 64.57.28.51 (64.57.28.51) 241.727ms 64.57.28.201 (64.57.28.201)
- 757.368ms 749.757ms
- 7 198.32.11.106 (198.32.11.106) 385.874ms 385.820ms 385.839ms
- 8 200.0.204.130 (200.0.204.130) 386.440ms 386.427ms 386.377ms
- 9 200.143.252.70 (200.143.252.70) 394.697ms 393.998ms 394.001ms
- 10 200.143.255.45 (200.143.255.45) 394.753ms 394.309ms 394.286ms
- 11 200.143.193.129 (200.143.193.129) 394.068ms 394.102ms 394.063ms

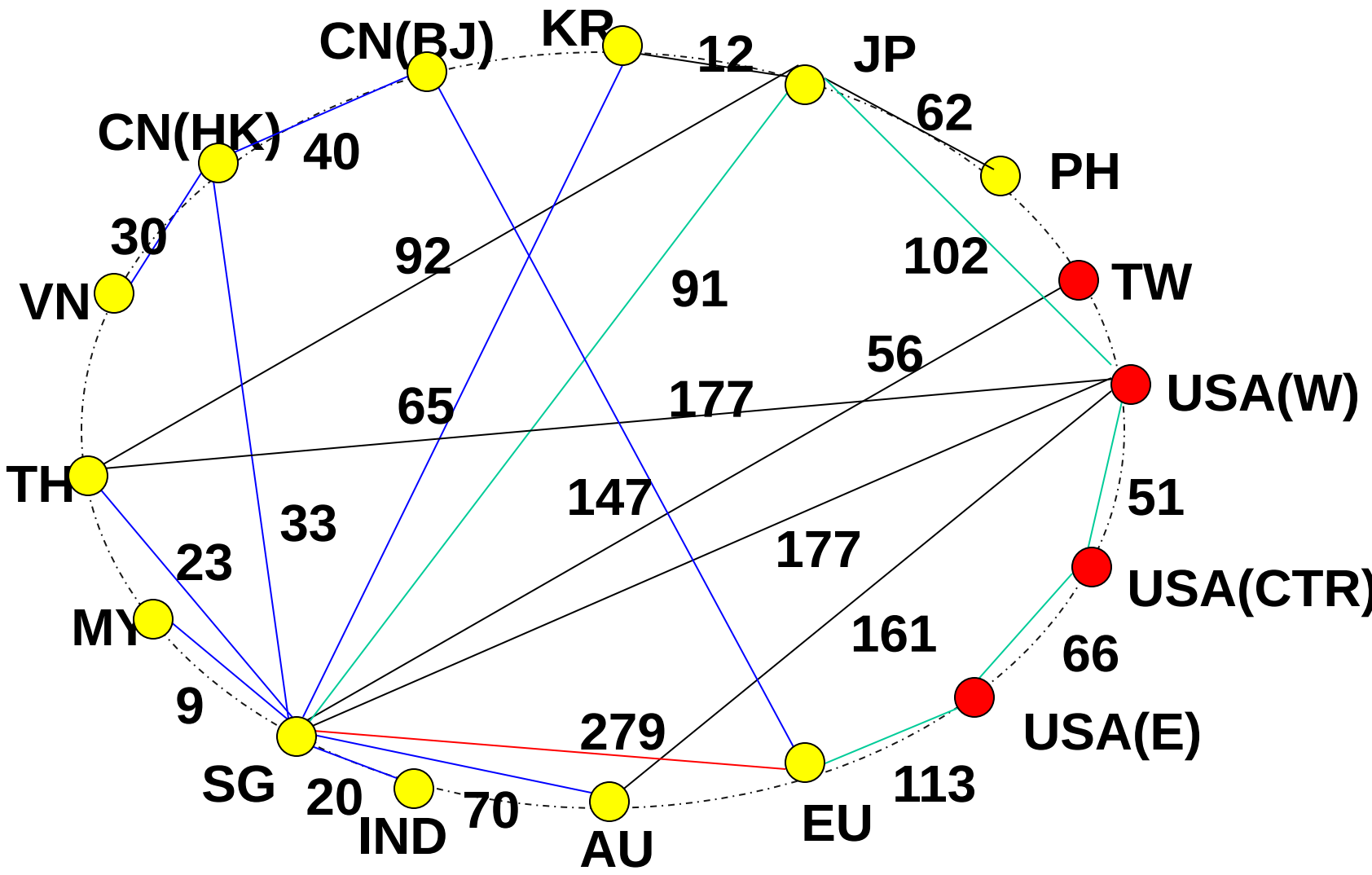
Tracert from Portugal to NTU via commercial ISP



tracert network tool



TEIN tracert network map



Created: 31 Oct 2006

Summary of IP