VM TCP/IP Routing - Part 1

Session V22

Alan Altmark
IBM Corporation

IBM Systems

Disclaimer

This presentation provides in-depth information on configuration of the routing components of VM TCP/IP.

References to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe on any of the intellectual property rights of IBM may be used instead. The evaluation and verification of operation in conjunction with other products, except those expressly designed by IBM, are the responsibility of the user.

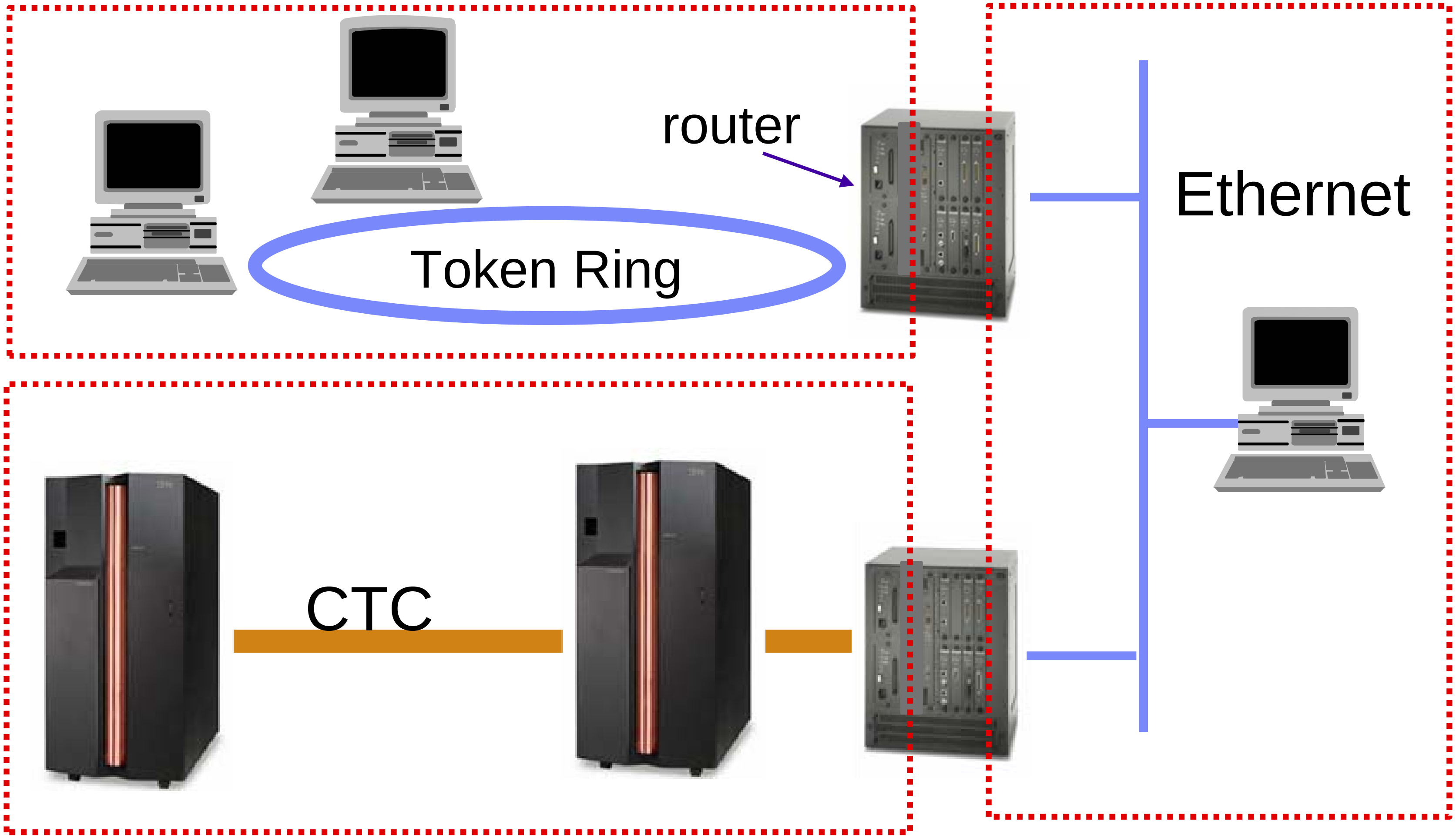
The following terms are trademarks of the International Business Machines Corporation in the United States or other countries or both: IBM IBM logo z/VM

Other company, product, and service names, which may be denoted by double asterisks (**), may be trademarks or service marks of others.

Agenda

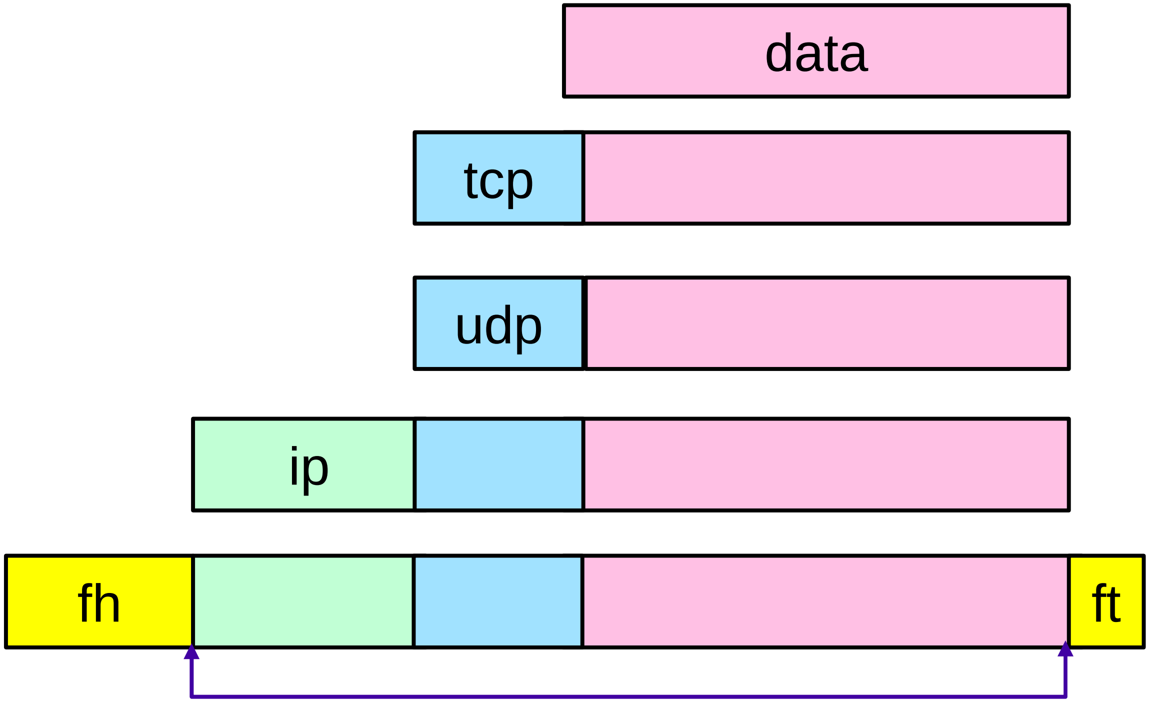
- Link-level communications
  - MAC frames
  - ARP
  - Proxy ARP
- IP Addressing
  - Classes
  - Subnets
- Routing basics

Terminology: LAN Segment



More Terminology

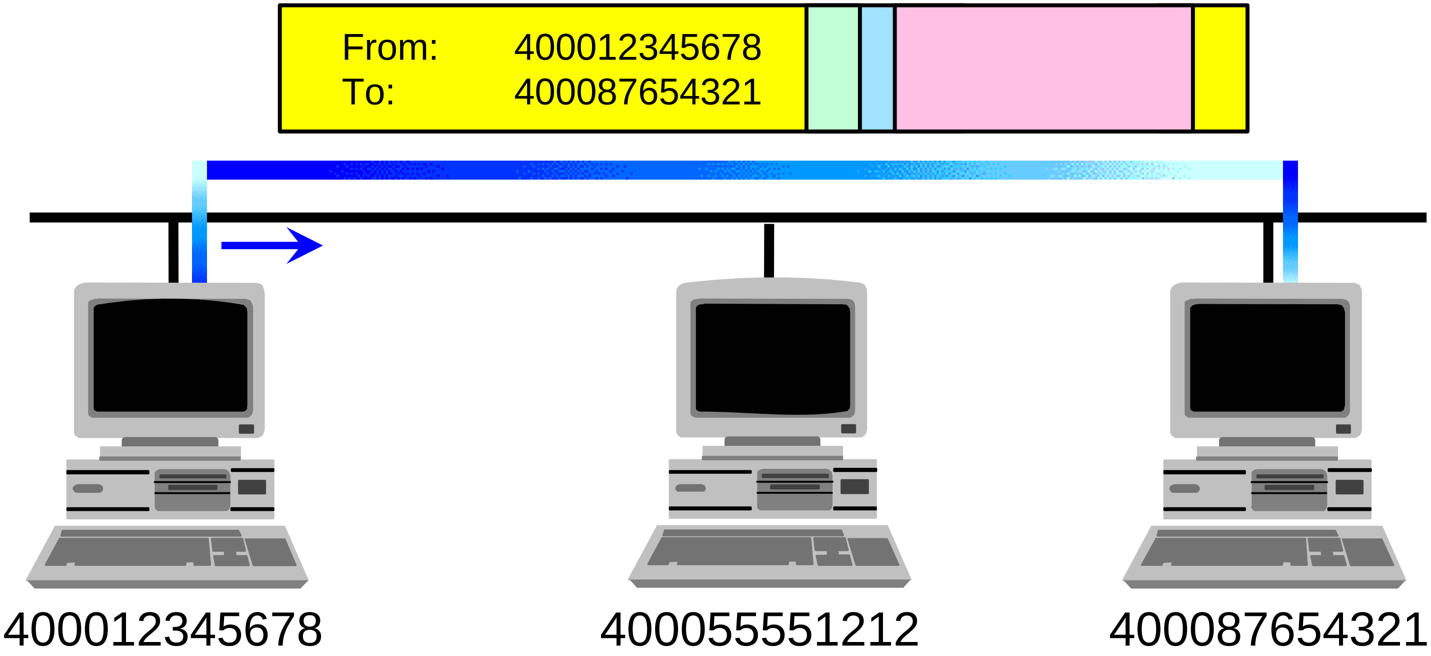
- Application data
- TCP Segment
- UDP Datagram
- IP Packet
- Link Frame
- MTU



Maximum **T**ransmission **U**nit

Link Level Communication - Unicast

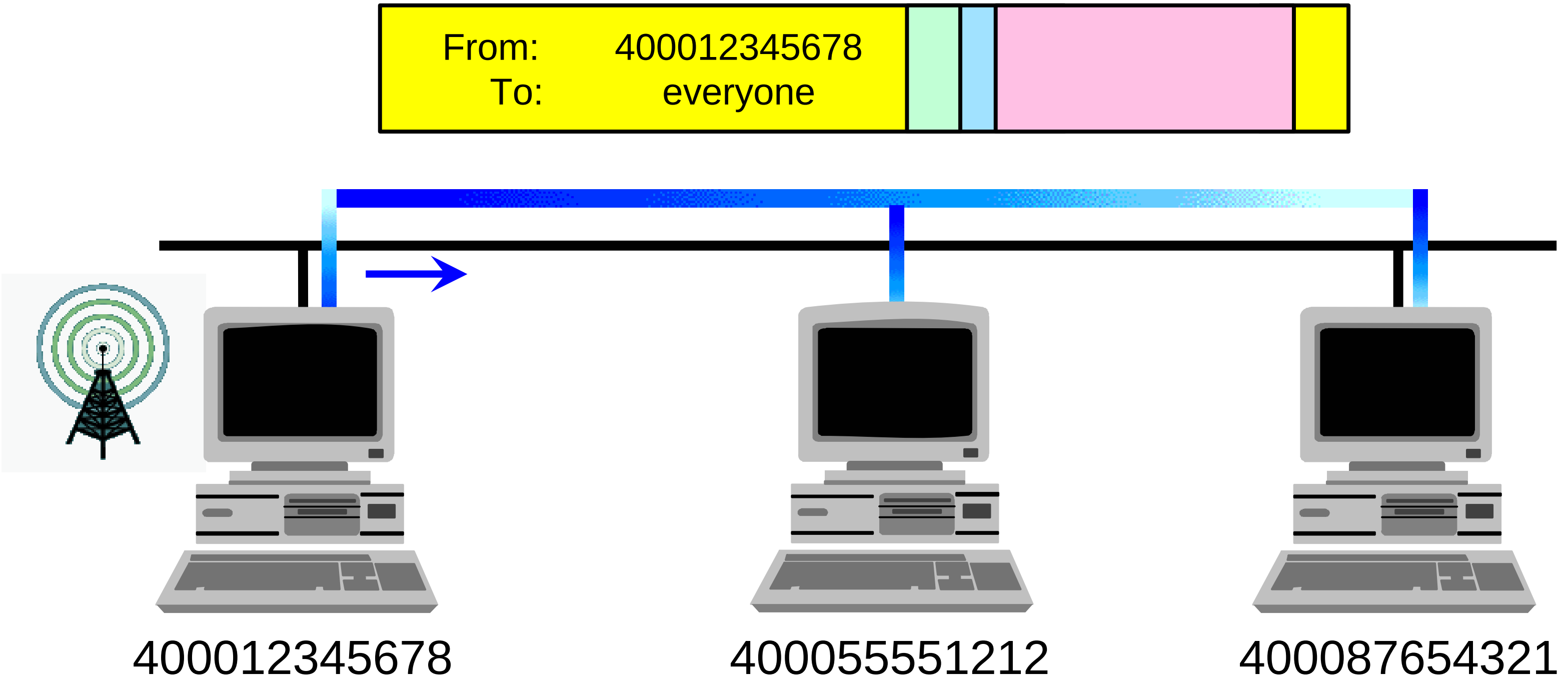
- Frames transmitted using Medium Access Control points and addresses



- Only addressed station picks up frame

Link Level Communication - Broadcast

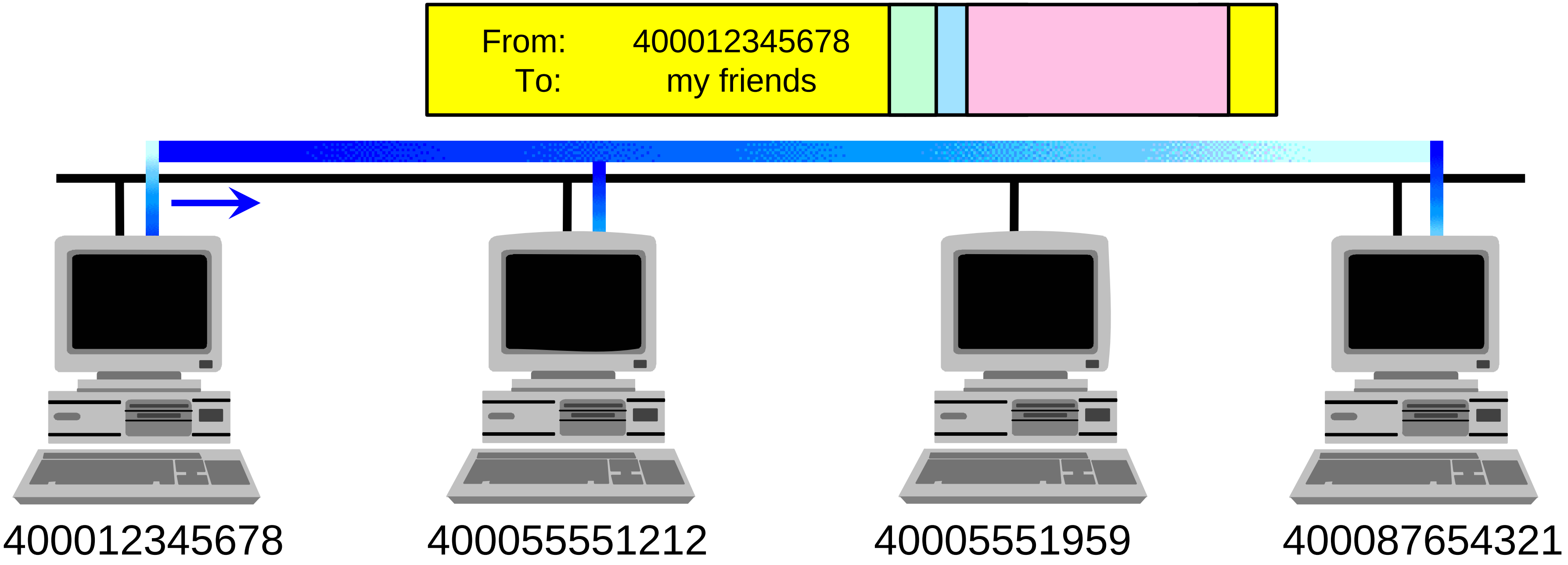
- Station can broadcast by using special format frame



- All stations will pick up frame

Link Level Communication - Multicast

- Station can reach all listening machines on LAN using a special multicast MAC address



- All stations registered for the multicast MAC address will pick up the frame

Converting a MAC address to an IP address

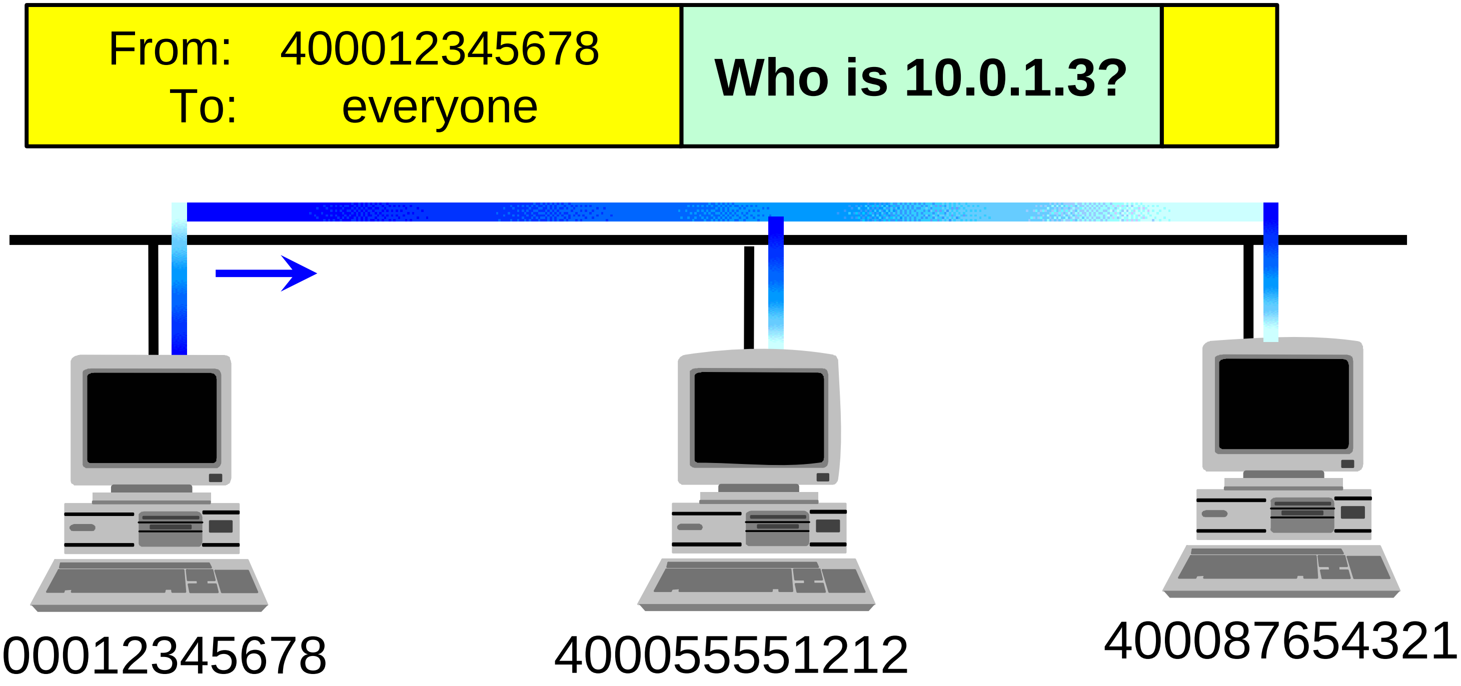
- IP hosts are managed and addressed using an IP address
- IP addresses are logical addresses, not physical

- So, how does TCP/IP convert an IP address to a physical MAC address?

- Answer: **Address Resolution Protocol (ARP)**

ARP Request

- Host broadcasts lookup on local LAN segment
  - Payload contains requested IP address

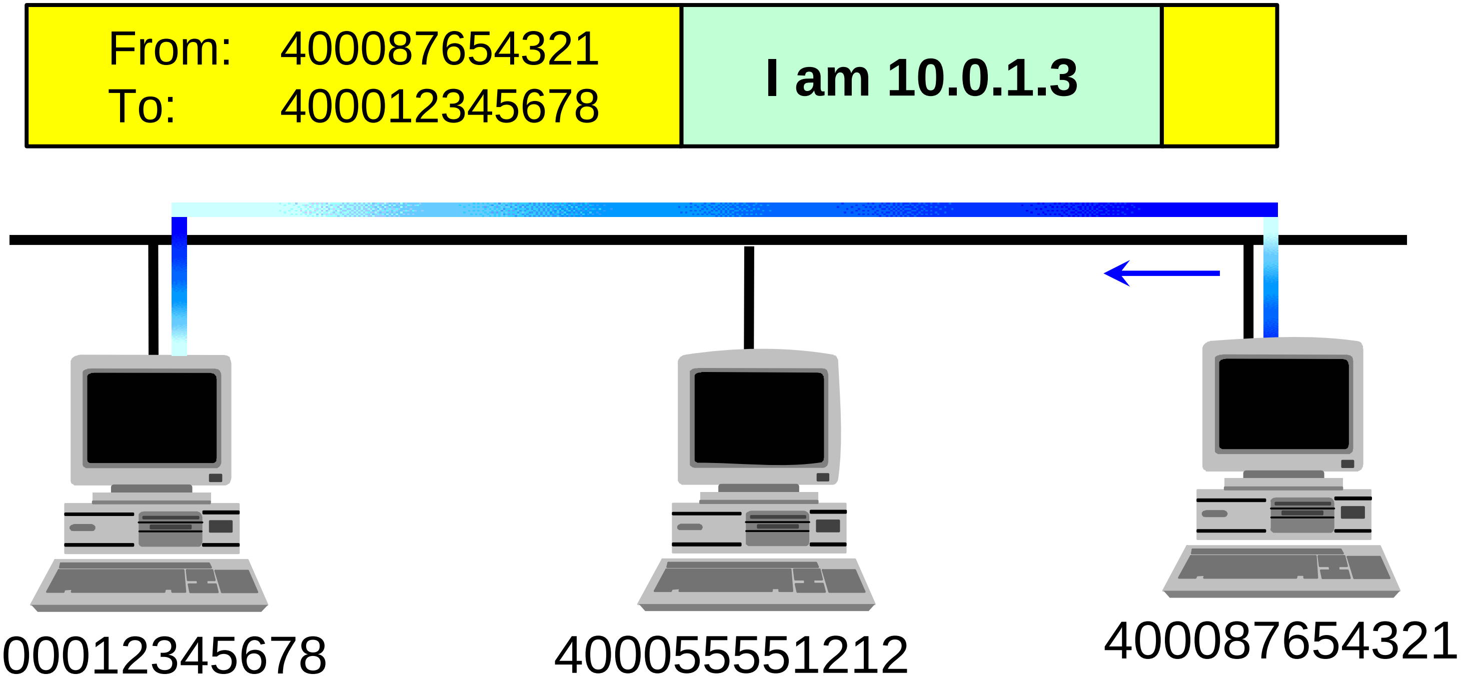


From: 400012345678
To: everyone

Who is 10.0.1.3?

ARP Response

- Owner of IP address responds with unicast response
  - What happens if two hosts have the same IP address?



From: 400087654321
To: 400012345678

I am 10.0.1.3

400012345678

400055551212

400087654321

ARP Cache

- Hosts maintain a cache of ARP responses to avoid ARP before sending each frame

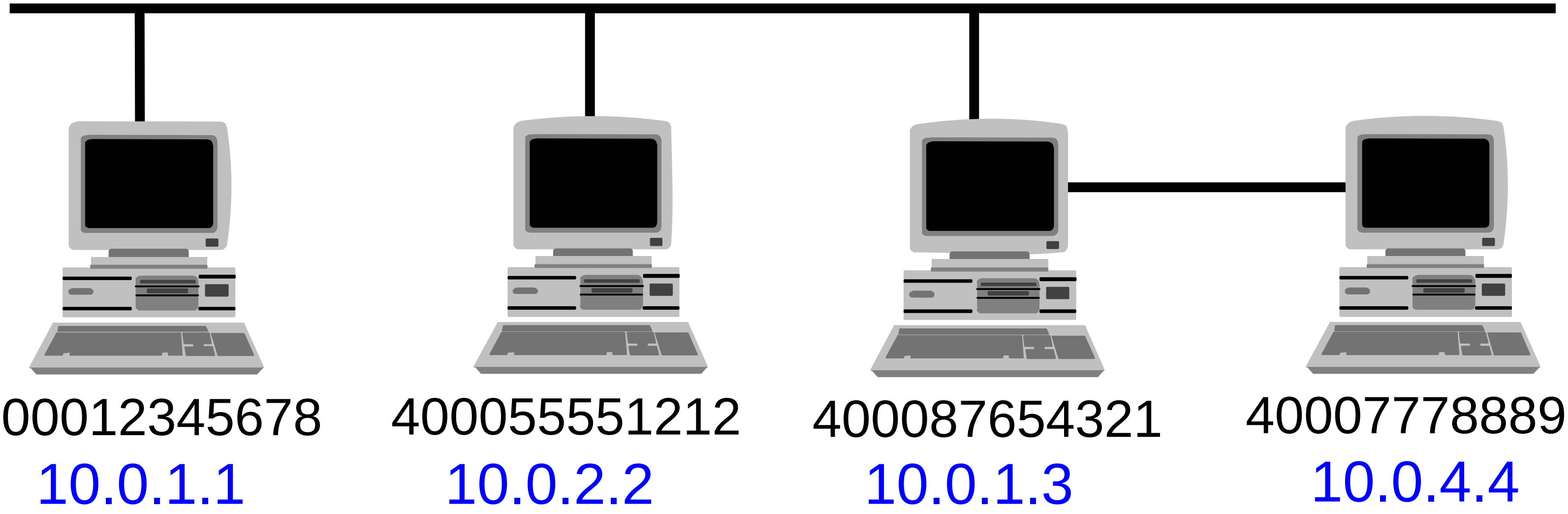
- ARP cache entries expire so that hosts can discover MAC address changes
  - New adapter
  - Different box with same IP address
    e.g. hot standby

Sample cache contents

10.0.1.1	400012345678	Timestamp
10.0.1.2	400055551212	Timestamp
10.0.1.3	400087654321	Timestamp

Proxy ARP: Bending the Rules

- In the following network configuration
  - Are 10.0.1.1 and 10.0.4.4 in the same subnet?
  - What would happen if 10.0.1.1 ARPs for 10.0.4.4?

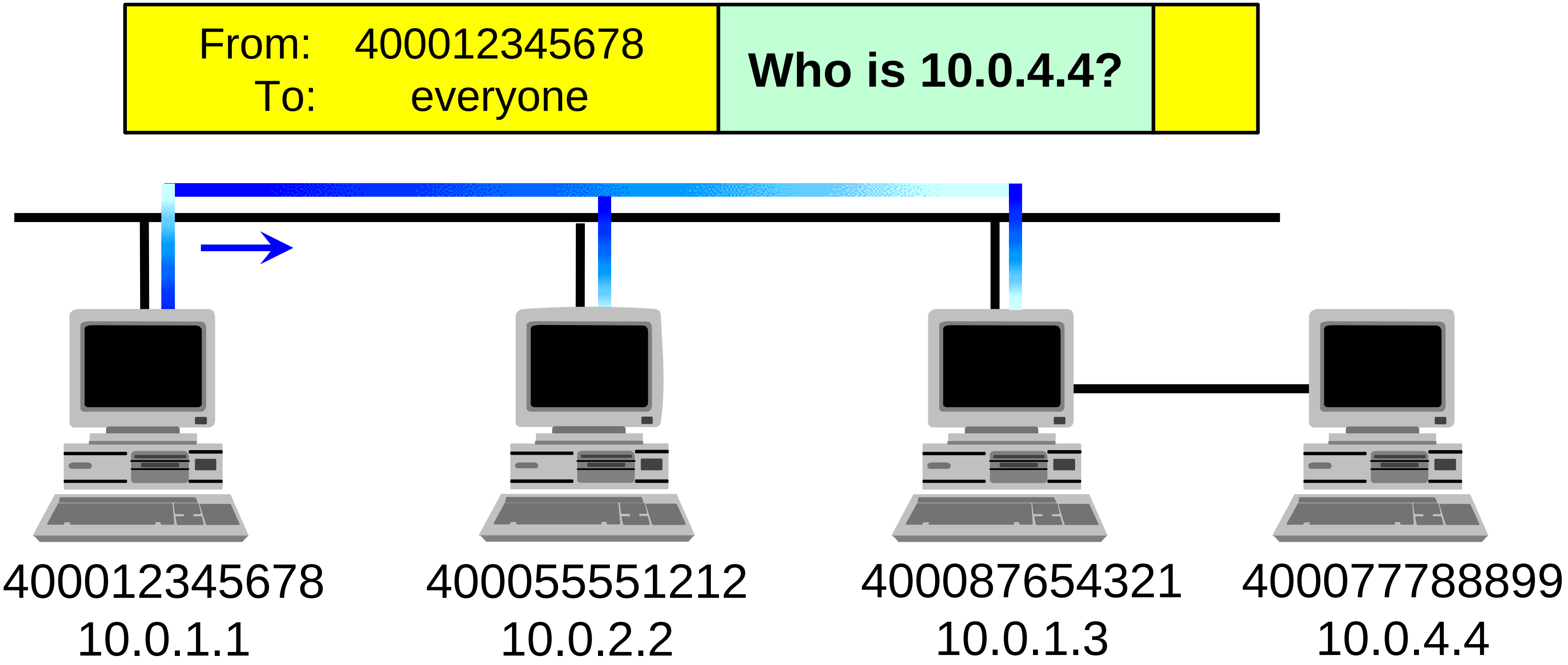


Subnet mask = 255.255.0.0

Surprise!

ARP Request

- 10.0.1.1 *assumes* that 10.0.4.4 is on the LAN because it is in the same subnet, so it ARPs



From: 400012345678
To: everyone

Who is 10.0.4.4?

400012345678
10.0.1.1

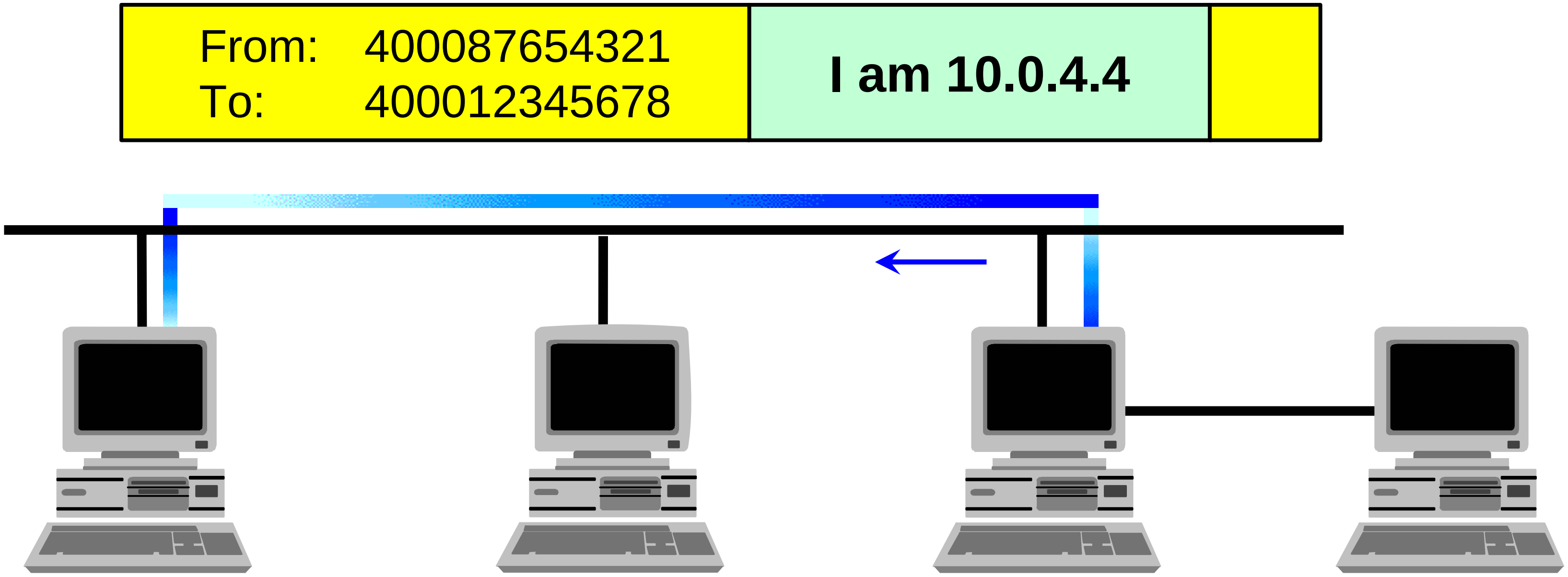
400055551212
10.0.2.2

400087654321
10.0.1.3

400077788899
10.0.4.4

Proxy ARP Response

- 10.0.1.3 pretends it is 10.0.4.4
  - “hidden router”



From: 400087654321
To: 400012345678

I am 10.0.4.4

400012345678
10.0.1.1

400055551212
10.0.2.2

400087654321
10.0.1.3

400077788899
10.0.4.4

Proxy ARP Configuration

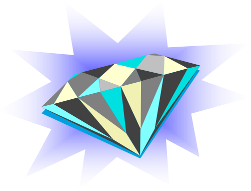
-
  ```
  AssortedParms
                 ProxyARP
  EndAssortedParms
  ```

- z/VM will respond on behalf of another host
  - Not controllable on a per-interface basis
  - HOST route entry required
  - Host must be same subnet as interface ARP arrives on

- Broadcast and multicast packets will not be forwarded

Local vs. Remote Hosts

- **Local** hosts are on same LAN segment and can be reached via ARP or proxy ARP
- **Remote** hosts must be reached through a local gateway or router
  - Each host has a default gateway defined to it
- Proxy ARP blurs the line
  - may provide SHORT-TERM alternative
  - does not solve all problems

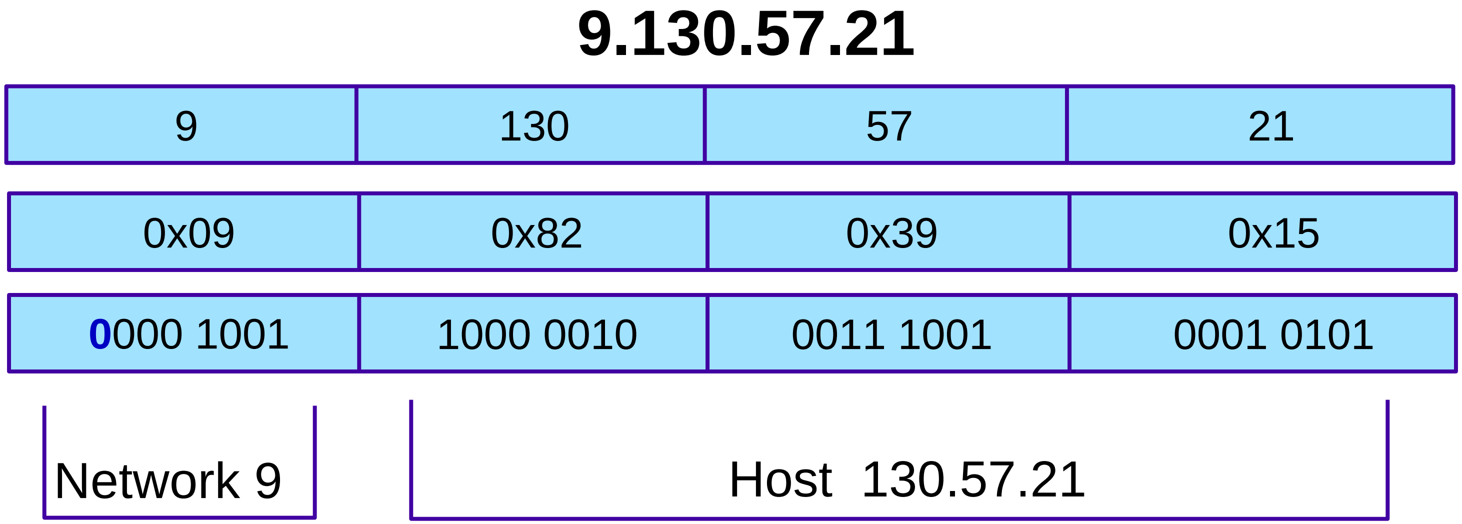


IPv4 Addressing

- 32-bit address, 4 *octets*
  - High-order bits identify *network*
  - Low-order bits identify *host* within network
  - Expressed as a.b.c.d
- Special values for network and host
  - All ones = "everyone"
  - All zeros = "me", "this", or "default"
- Address space divided into classes
  - For convenience only
  - Some defaults are based on class

IPv4 Addressing: Class A

- Networks: 0 to 127
  Total: 128 networks



9.130.57.21

9	130	57	21
0x09	0x82	0x39	0x15
0000 1001	1000 0010	0011 1001	0001 0101

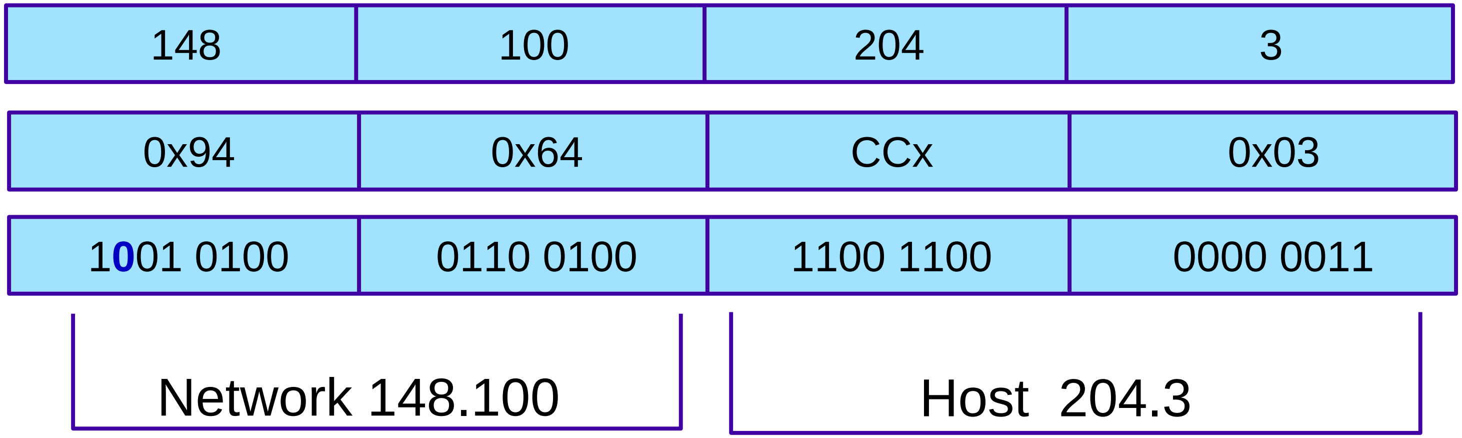
Network 9

Host 130.57.21

IPv4 Addressing: Class B

- Networks: 128.0 to 191.255
  Total: 16 384 networks

148.100.204.3



148	100	204	3
0x94	0x64	CCx	0x03
1001 0100	0110 0100	1100 1100	0000 0011

Network 148.100

Host 204.3

IPv4 Addressing: Class C

- Networks: 192.0.0 to 223.255.255
  Total: 2 097 152 networks

200.14.64.191

200	14	64	191
0xC8	0x0E	0x40	0xBF
11**0**0 1000	0000 1110	0100 0000	1011 1111

Network 200.14.64

Host 191

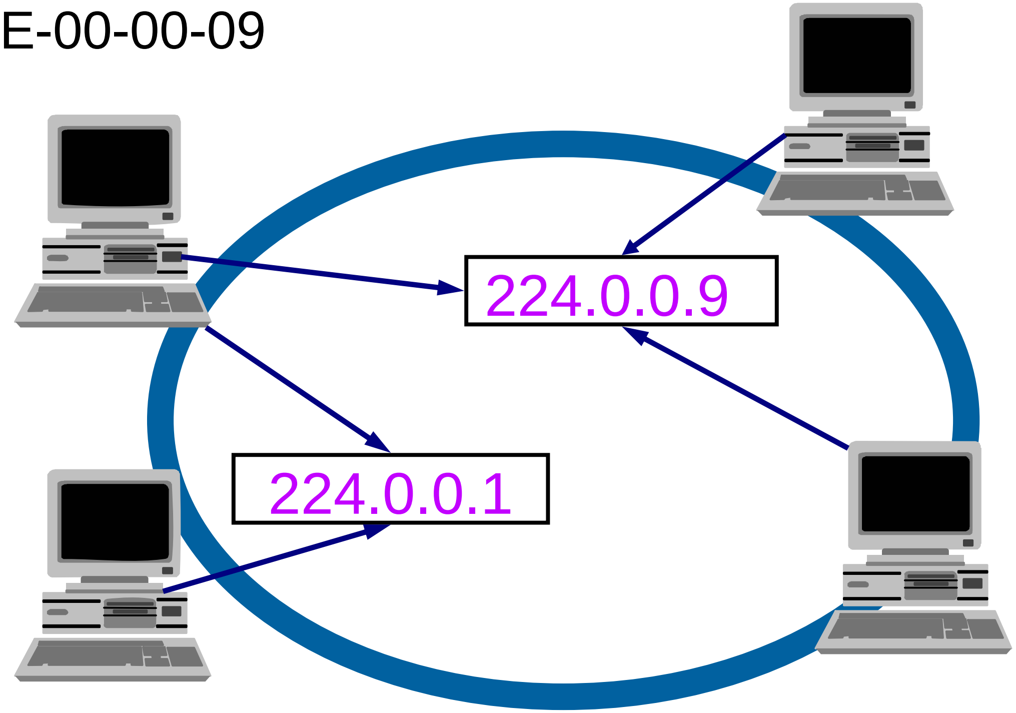
IPv4 Addressing: Class D Multicast

- 224.0.0.0 to 239.255.255.255
- provides 28-bit multicast group id
  - low-order 23 bits used in ethernet address 01-00-5E-00-00-00
  - E.g. 224.0.0.9 = 01-00-5E-00-00-09

- Hardware facility

- Limited broadcast reduces unnecessary processing by uninterested parties

- Used by RIPv2 and OSPF routers, and IPv6

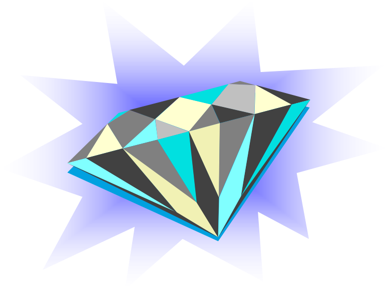


Subnetting

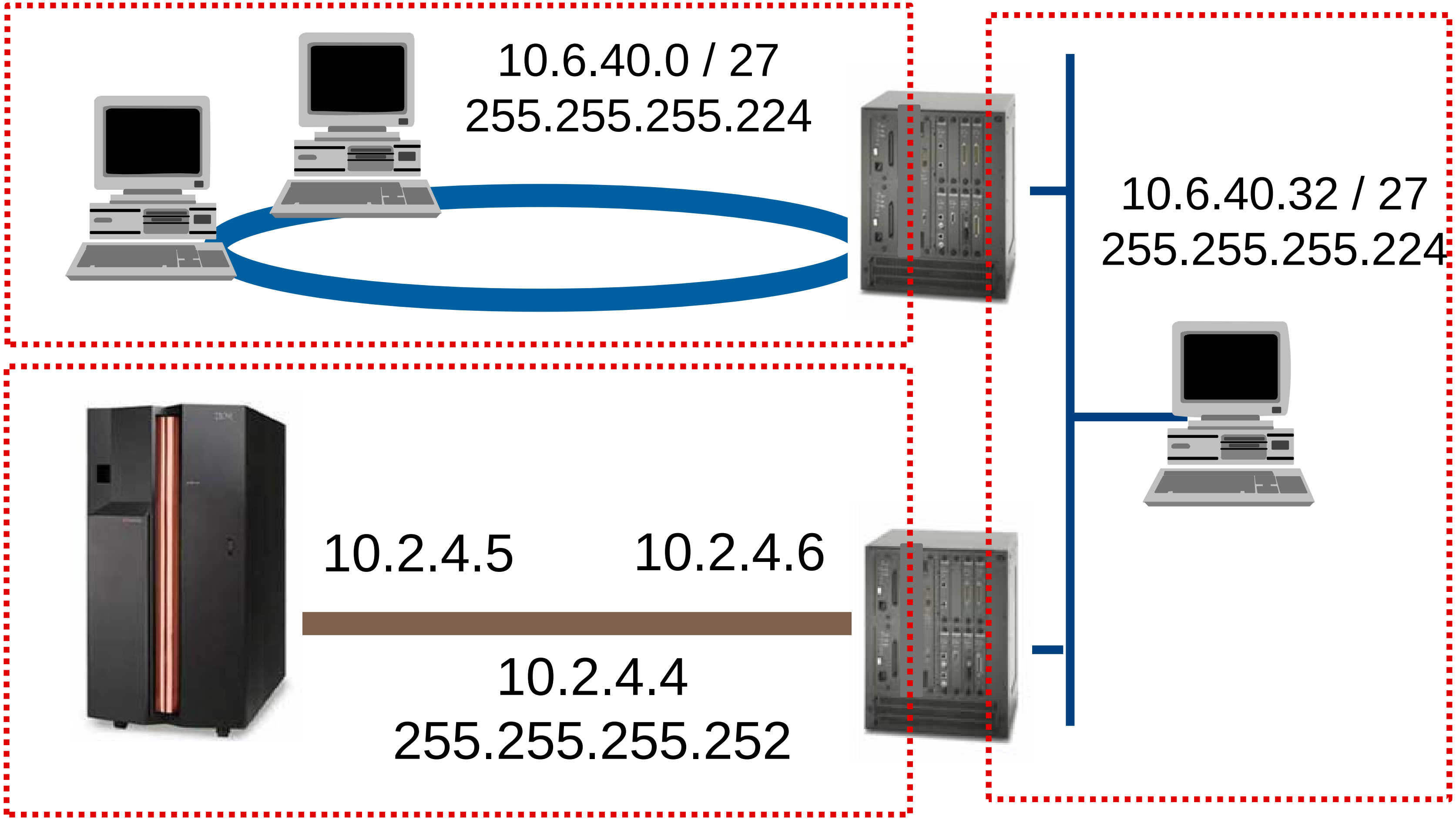
- Class A and B networks provide for 16M and 64K hosts, respectively
- LAN segments do not contain anywhere near that many hosts
- Divide host id portion of address into manageable groups called *subnets*
- In general, classes are used for convenience
  - CIDR - Classless Internet Domain Routing
  - Everything uses subnet masks

Subnetting

- Hosts that are members of the same subnet are considered to be in the same LAN segment
- Point-to-point is a "LAN segment" with exactly two hosts
- Multiple subnets may share same LAN segment
  - a.k.a "multinet"



Subnet = LAN Segment



Subnetting

- The subnet mask defines which bits of the host id are used for the subnet number

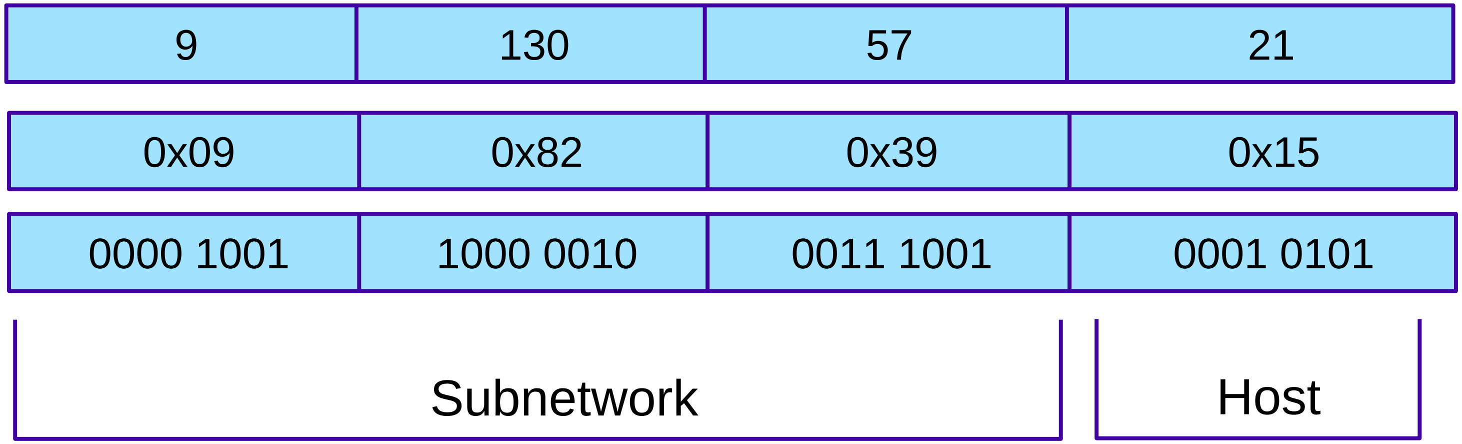
- Subnet number = bitand(address, mask)

  Perform logical AND of destination address and subnet mask to get subnet number

  bitand(9.130.3.157, 255.255.255.240) = 9.130.3.144

IPv4 Subnet Addressing

Subnet mask = 255.255.255.0 (/24)
IP address = 9.130.57.21

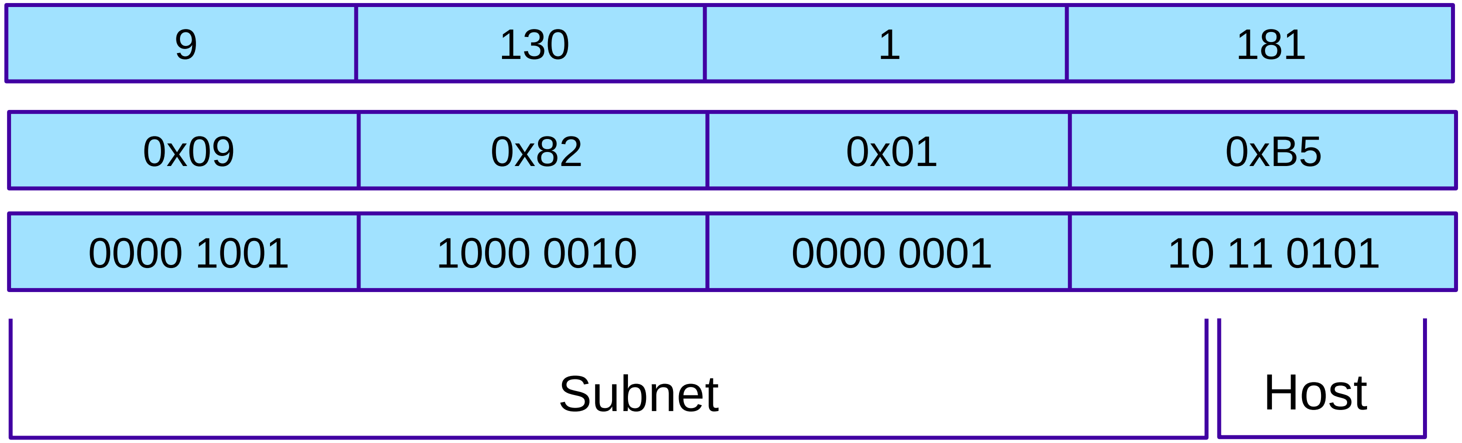


9	130	57	21
0x09	0x82	0x39	0x15
0000 1001	1000 0010	0011 1001	0001 0101

Subnet = 9.130.57.0

IPv4 Subnet Addressing

Subnet mask = 255.255.255.192 (/26)
IP address = 9.130.1.181



9	130	1	181
0x09	0x82	0x01	0xB5
0000 1001	1000 0010	0000 0001	10 11 0101

Subnet value = 9.130.1.128
(How was this determined?)

IPv4 Subnet Addressing

Subnet mask = 255.255.255.192 (/26)
IP address = 9.130.1.181

	0000 1001	1000 0010	0000 0001	1011 0101
&	1111 1111	1111 1111	1111 1111	1100 0000
=	0000 1001	1000 0010	0000 0001	1000 0000
=	9	130	1	128

Subnet = 9.130.1.128
Host = 53 (0x35)

0011 0101

Remaining bits are host number
(cannot be all 1's or all 0's!)

IPv4 Addressing Quick Reference

Class	First octet	Network
A	0-127	a.0.0.0
B	128-191	a.b.0.0
C	192-223	a.b.c.0
D	224-239	n/a

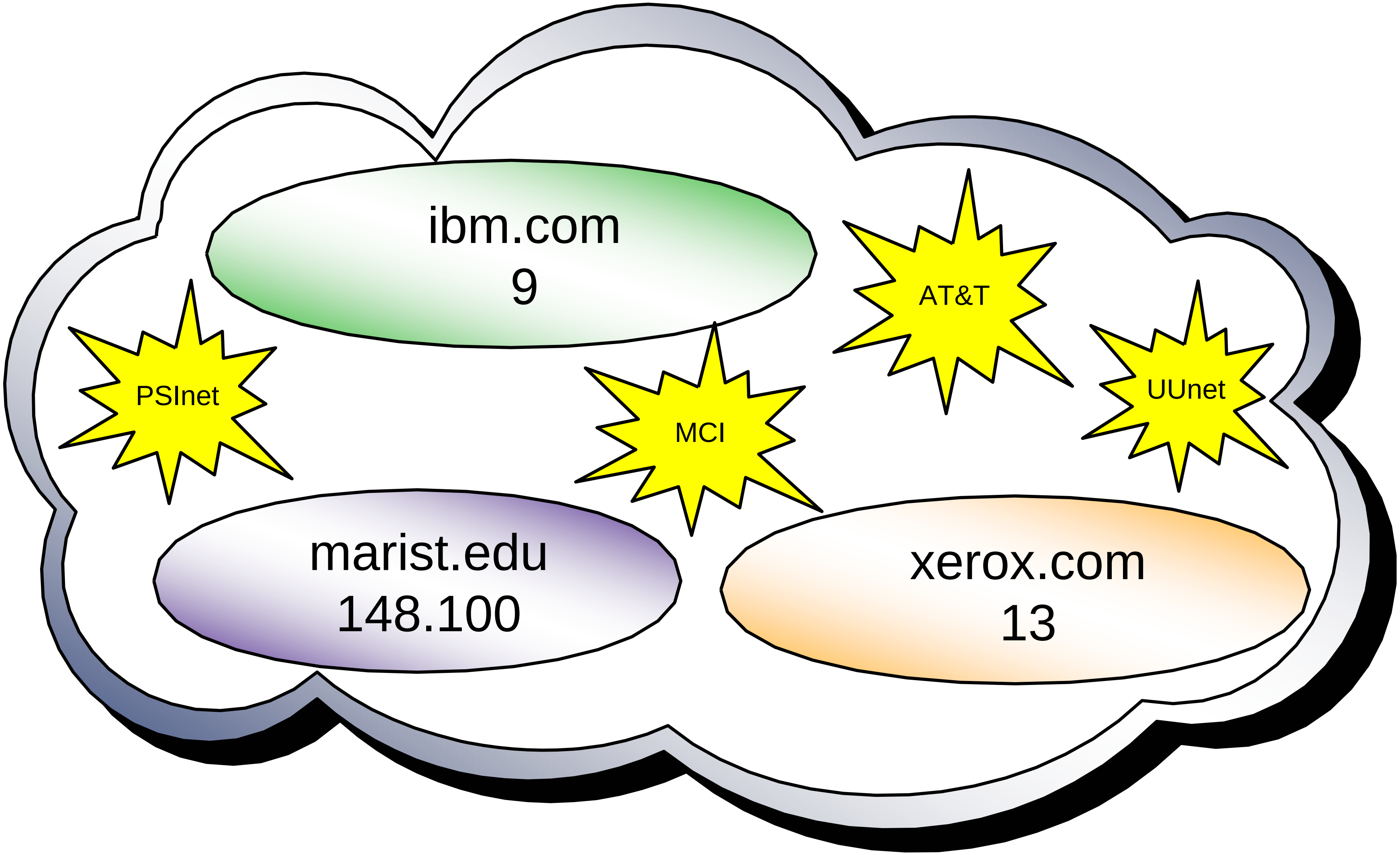
Mask size	Last octet	binary	subnetwork	# hosts
/25	128	1000 0000	**2:** 0 128	126
/26	192	1100 0000	**4:** 0 64 128 192	62
/27	224	1110 0000	**8:** 0 32 64 96 128 160 192 224	30
/28	240	1111 0000	**16:** 0 16 32 48 64 80 96 112 128 144 160 176 192 208 224 240	14
/29	248	1111 1000	**32:** 0 8 16 24 32 40 48 56 64 72 80 88 96 104 112 120 128 136 144 152 160 168 176 184 192 200 208 216 224 232 240 248	6
/30	252	1111 1100	**64:** 0 4 8 16 20 24 28 32 36 ...	2

Special IPv4 Addresses

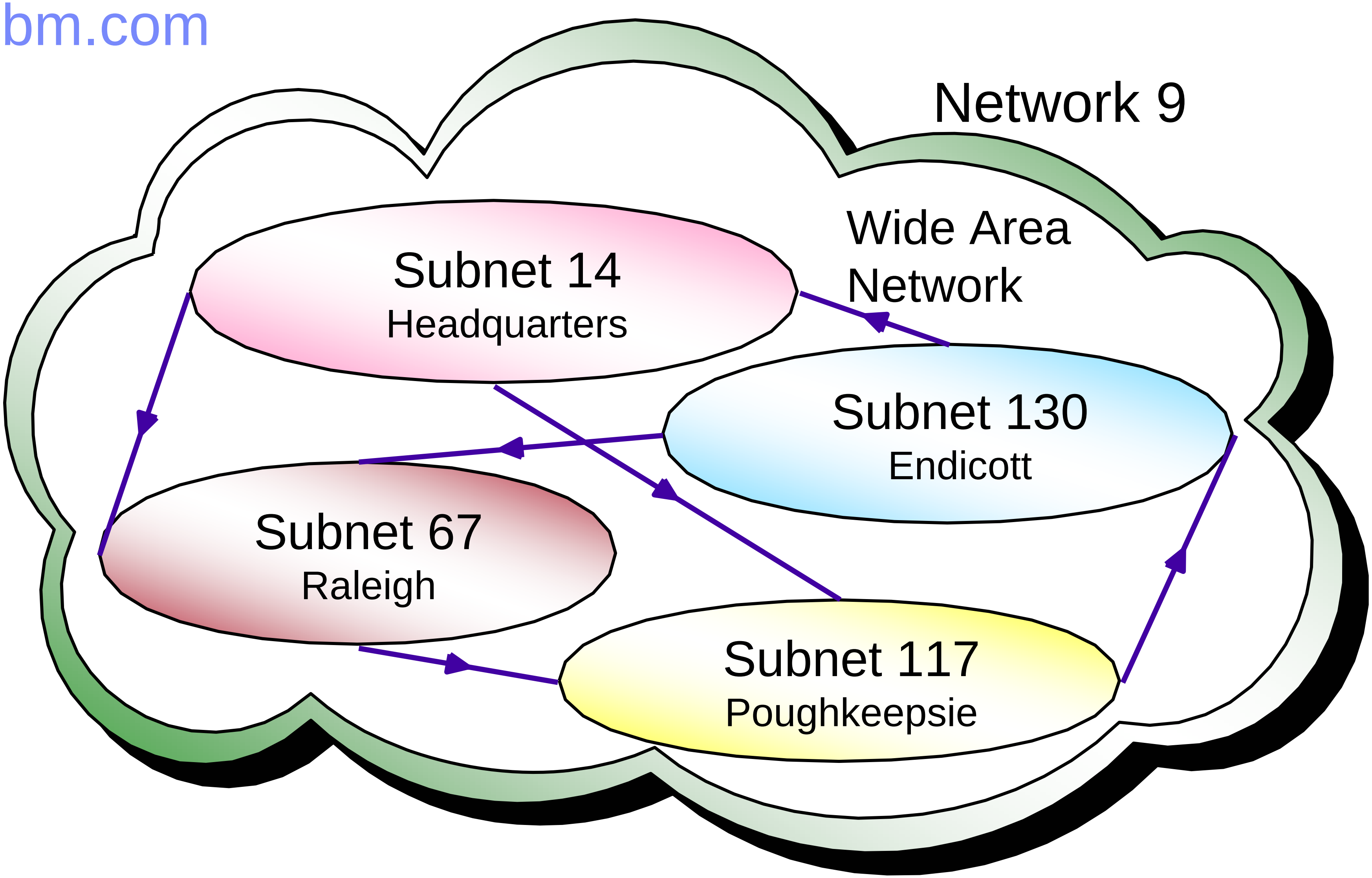
net ID	**subnet ID**	**host ID**	**Source**	**Destination**	**Description**
0		0	yes	no	this host on this net
0		*hostid*	yes	no	specific host on this net
127		*any*	yes	yes	Loopback
-1		-1	no	yes	local media broadcast
netid		-1	no	yes	network-directed broadcast
netid	*subnetid*	-1	no	yes	subnet-directed broadcast
netid	-1	-1	no	ok	all-subnets-directed broadcast

Local broadcasts are not bridged or routed to other LAN segments

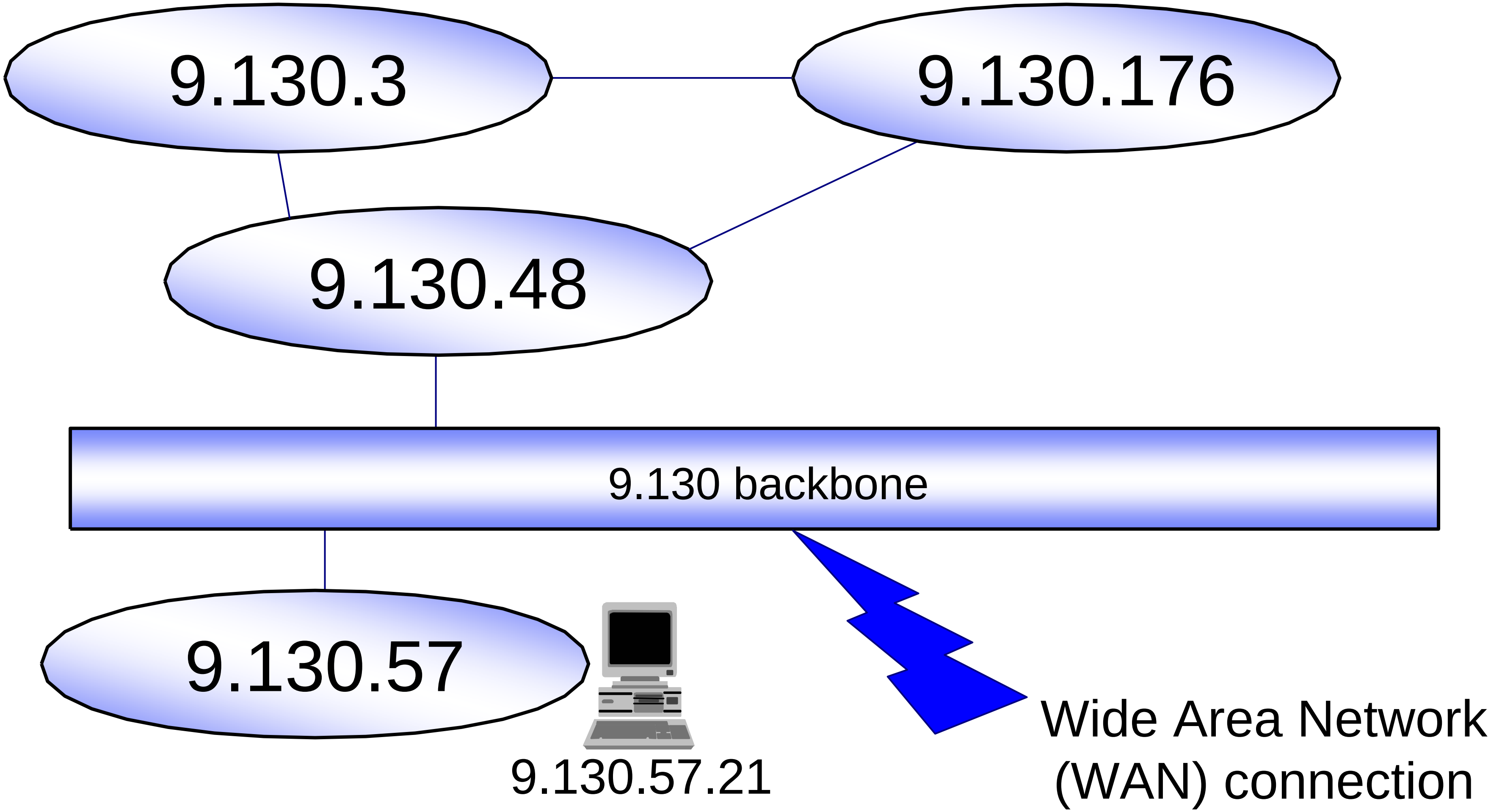
Networks on the Internet



ibm.com


Network 9
Wide Area Network
Subnet 14
Headquarters
Subnet 130
Endicott
Subnet 67
Raleigh
Subnet 117
Poughkeepsie

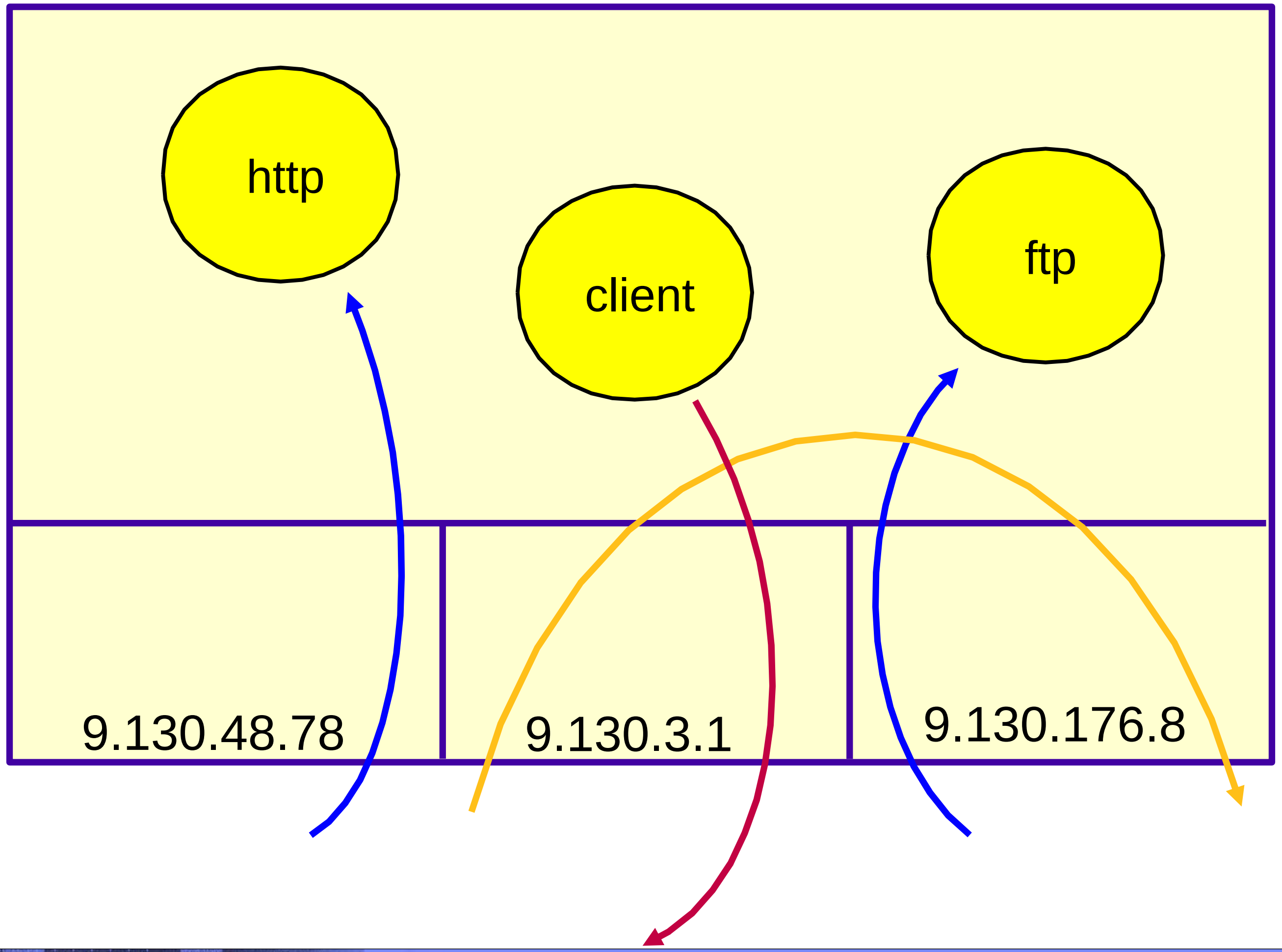
endicott.ibm.com



IP Packet Routing

- Occurs whenever an IP packet is received or sent by a host
  - Sometimes trivial - Only one possible route
  - Sometimes complex - Multi-homed host
- Like a game of "Hot Potato"
  - If not mine, make it someone else's problem ASAP!
  - Logic:
    - If it's for me, kick it upstairs
    - If it's for a host on a network to which I'm connected, send it (point to point) or ARP (LAN)
    - If for some other network, forward to someone else
    - Otherwise, drop it

Multi-homed Host



Routing

- The magic is in selecting the right host in order to reach some other network
- Failing to follow IP addressing rules regarding subnets and LANs result in "host unreachable" or timeouts.
- Describing the local network topology to your system involves learning arcane specification rules
- You will be considered wise and learned!

The Adventure Continues…

- Stay tuned for Routing - Part 2
  Don't touch that dial!

- We'll get into VM TCP/IP host configuration specifics
  - Static routing
  - Dynamic routing
  - VIPA - Virtual IP Addressing
  - Virtual Switching

Read More About It…

z/VM TCP/IP Planning and Customization	SC24-6019
TCP/IP Illustrated, Vol. 1 Addison Wesley	W. Richard Stevens ISBN 0-201-63346-9
Internetworking with TCP/IP Prentice Hall	Douglas P. Comer ISBN 0-13-216987-8

Contact Information

By e-mail:	Alan_Altmark@us.ibm.com
In person:	USA 607.429.3323
On the Web:	http://ibm.com/vm/devpages/altmarka
Mailing lists:	IBMTCP-L@vm.marist.edu VMESA-L@listserv.uark.edu LINUX-390@vm.marist.edu http://ibm.com/vm/techinfo/listserv.html