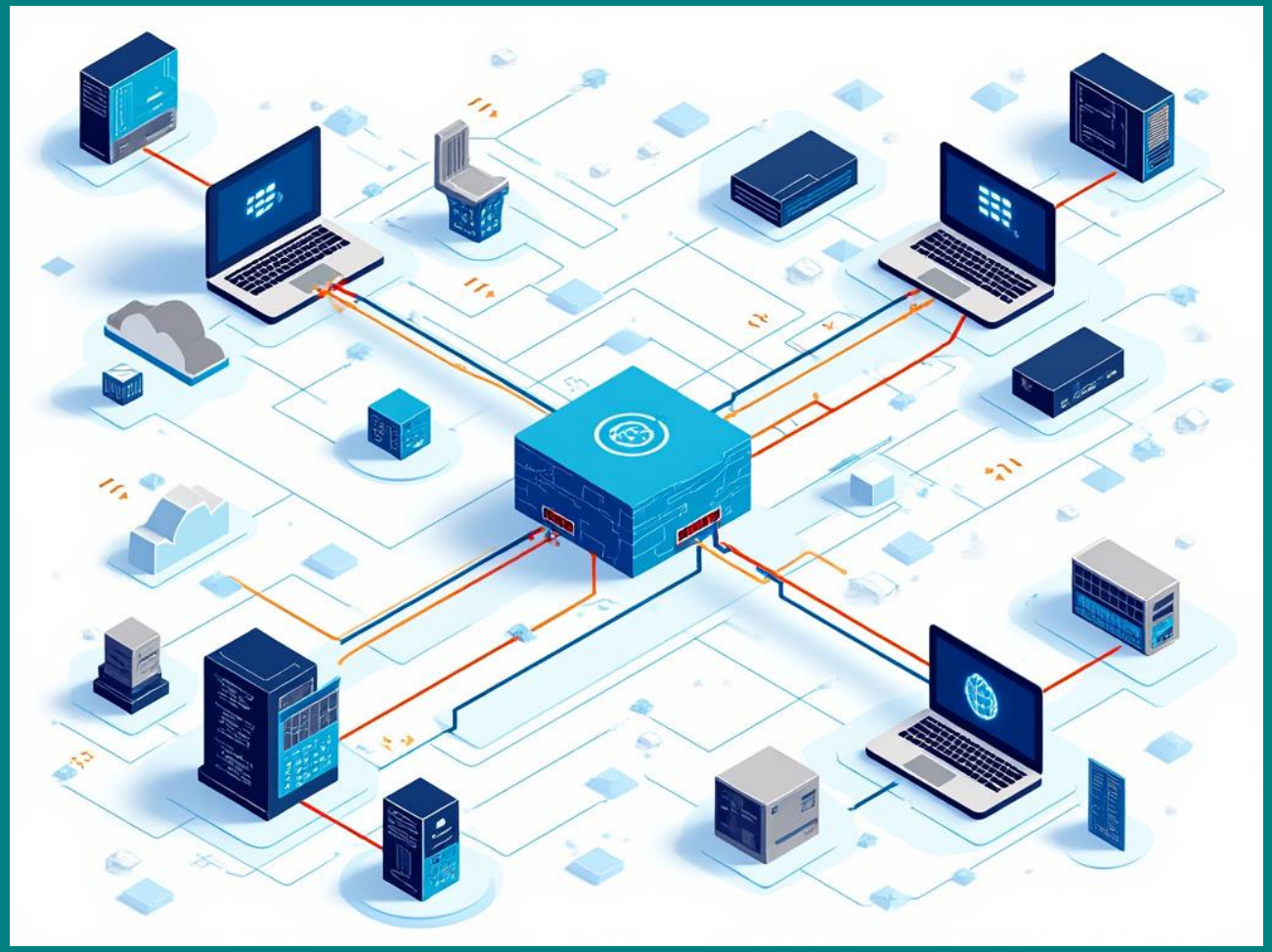
# TCP/IPv4 Network Training



# HISTORICAL BACKGROUND

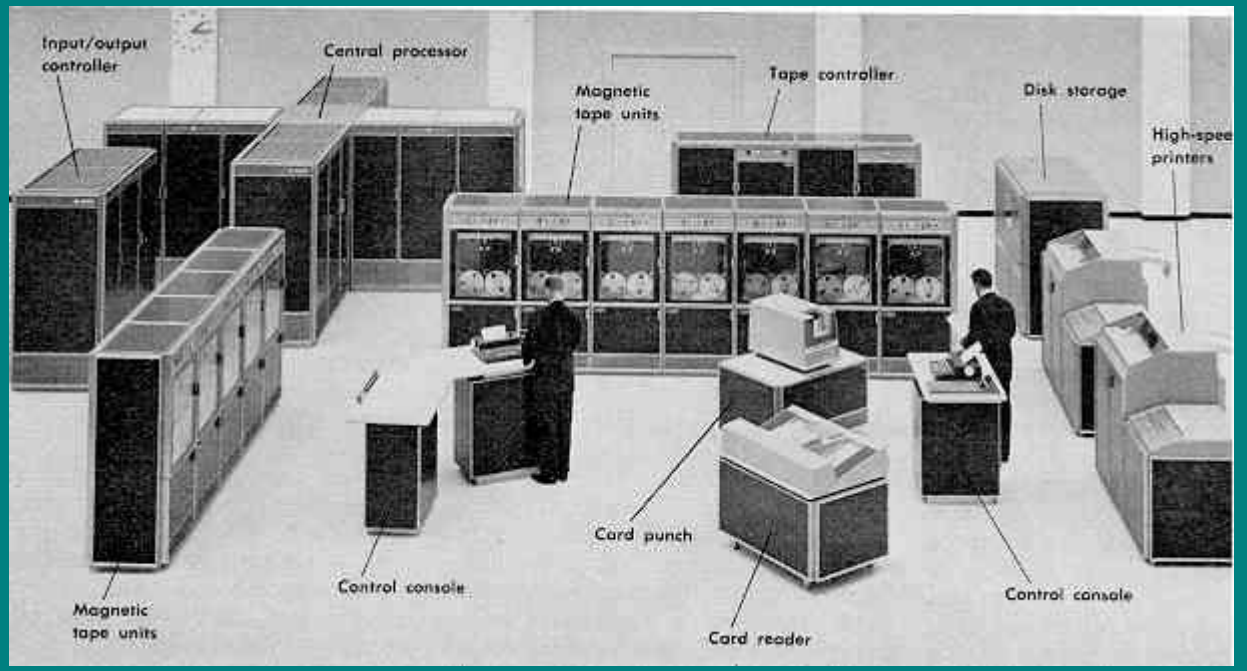
# The myth why the Internet was born

- The myth: *„The Army puts out a bid on a computer and DEC wins the bid. The Air Force puts out a bid and IBM wins. The Navy bid is won by Unisys. Then the President decides to invade Grenada, and the armed forces discover that their computers cannot talk to each other. The DOD must build a 'network' out of systems each of which, by law, was delivered by the lowest bidder on a single contract."*
  - Source: http://www.yale.edu/pclt/COMM/TCPIP.HTM
  - Note: The invasion of Grenada, a Caribbean island nation north of Venezuela, took place in 1983 under President Reagan. The military code-name was *„Operation Urgent Fury"* and the operation restored constitutional government in Grenada

# The role of (D)ARPA

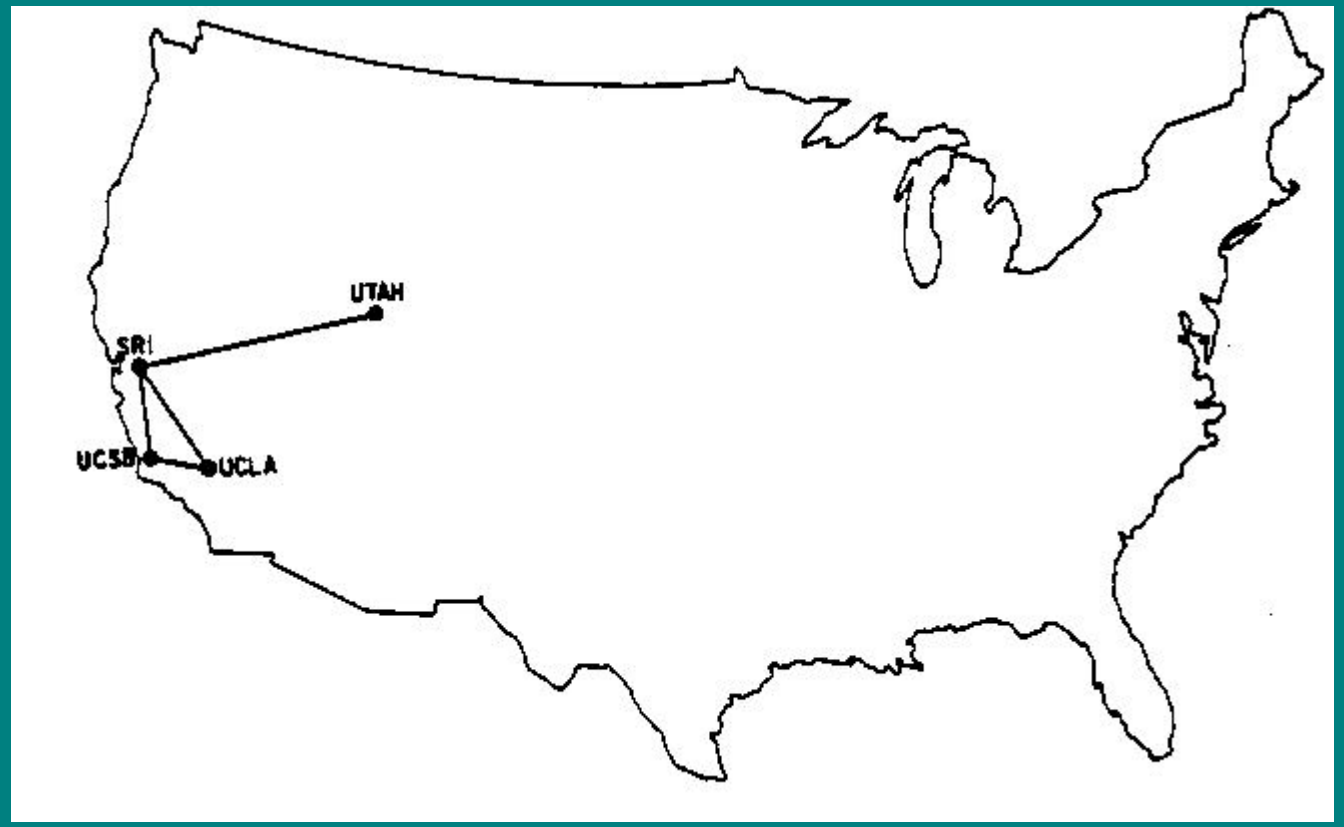
- DARPA or ARPA (“United States Defense Advanced Research Projects Agency”) developed secret systems and weapons during the Cold War and ran an own research network
- Development of the ARPANET, “the first Internet”, began in 1969
- The ARPANET was used to connect the few computers of the time that were geographically spread all over the United States; those computers were big data-center-sized mainframes, not small PCs
- The ARPANET was decommissioned in 1990. The Internet was commercialized in 1995

# A computer of the time

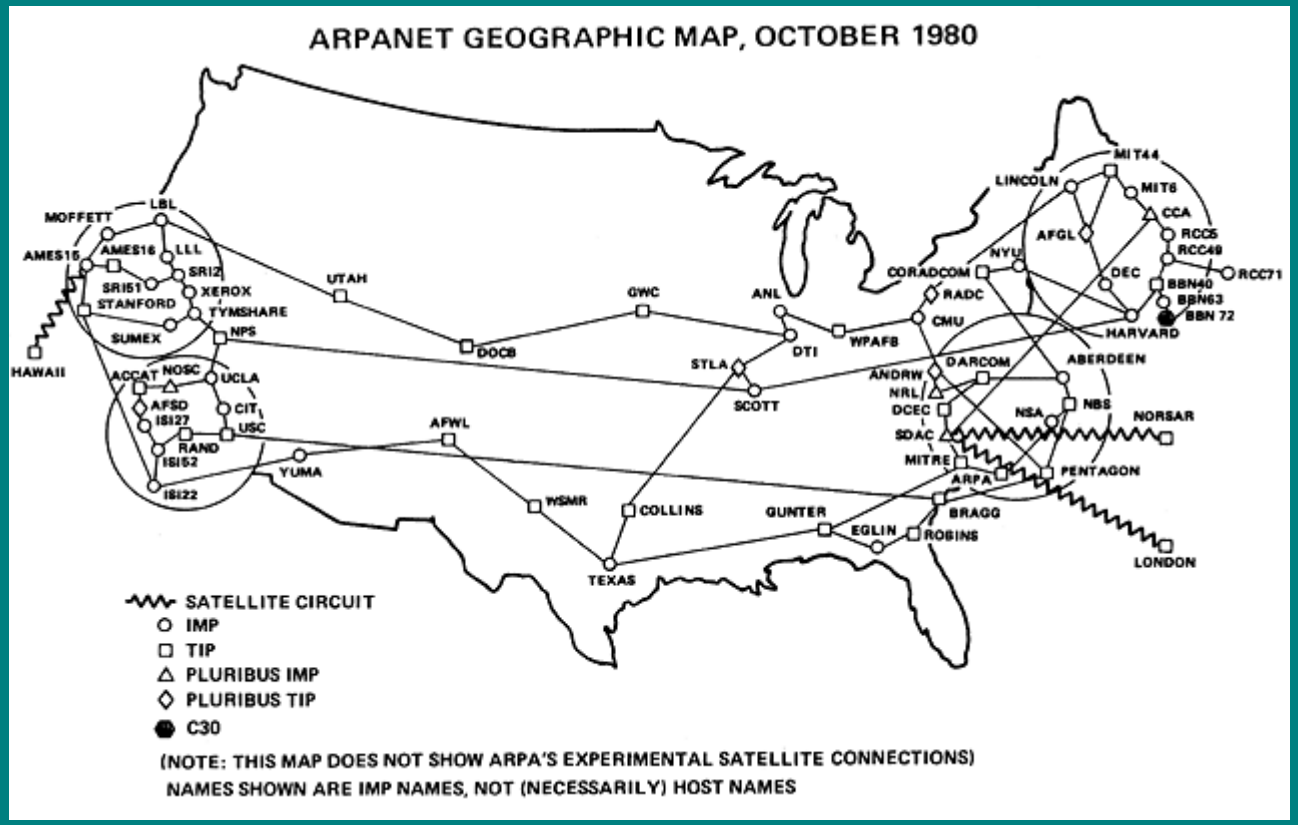



Not a PC, but still a computer

# ARPANET Geography in 1969



# ARPANET Geography in 1980



# The OSI Layer Model and TCP/IP

# The network language: TCP/IP

- The TCP/IP protocol family was developed as a part of the DARPA's research network
- TCP/IP stands for "Transmission Control Protocol/Internet Protocol"
- TCP/IP uses a four layer model, which basically is a simplified version of the 7-Layer Open Systems Interconnection ("OSI") model
- The TCP/IP standard and related protocols are defined in so-called RFCs ("Request For Comments")
- RFC 675 for the TCP protocol was published in December 1974; first use of the name and word "Internet" in this RFC
- IETF ("Internet Engineering Task Force") curates the RFCs: http://www.ietf.org/rfc.html

# The OSI 7-layers reference model

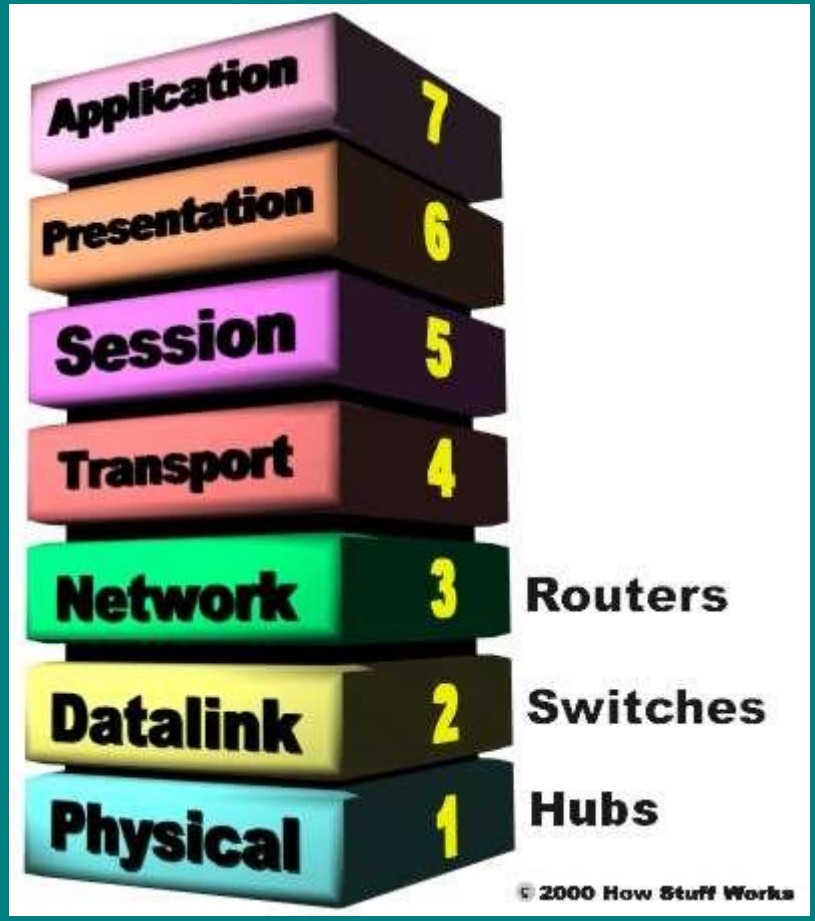
# The OSI layers 1-3

- Layer 1: Physical - This is the level of the actual hardware
- Layer 2: Data - In this layer, the appropriate physical protocol is assigned to the data. Also, the type of network and the packet sequencing is defined
- Layer 3: Network - The way that the data will be sent to the recipient device is determined in this layer. Logical protocols, routing and addressing are handled here

# The OSI layers 4-7

- Layer 4: Transport - This layer maintains flow control of data and provides for error checking and recovery of data between the devices
- Layer 5: Session - Layer 5 establishes, maintains and ends communication with the receiving device.
- Layer 6: Presentation - Layer 6 takes the data provided by the Application layer and converts it into a standard format that the other layers can understand
- Layer 7: Application - This is the layer that actually interacts with the operating system or application whenever the user chooses to transfer files, read messages or perform other network-related activities

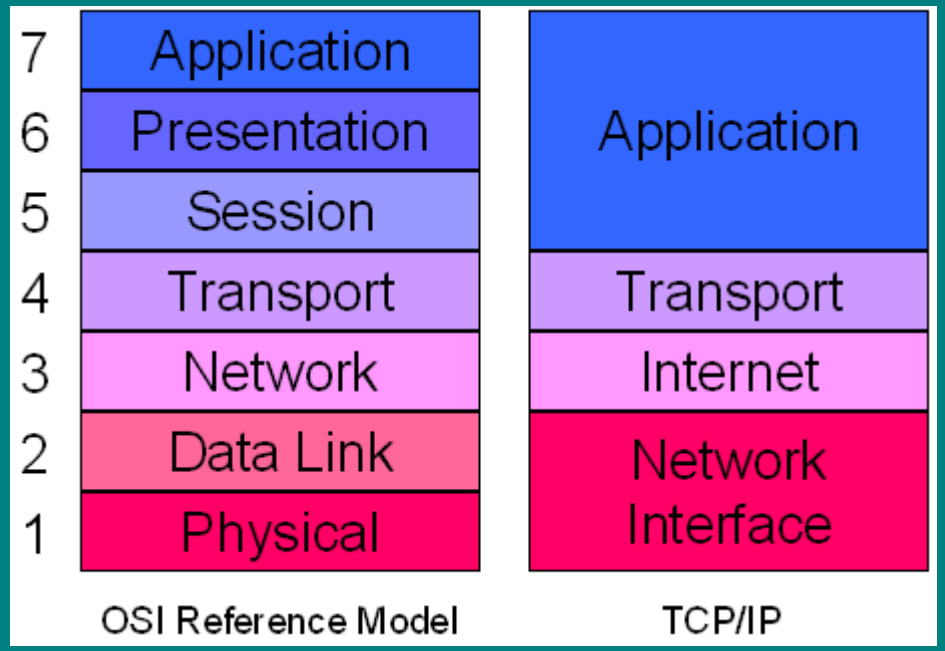
# On which OSI layer works our equipment?



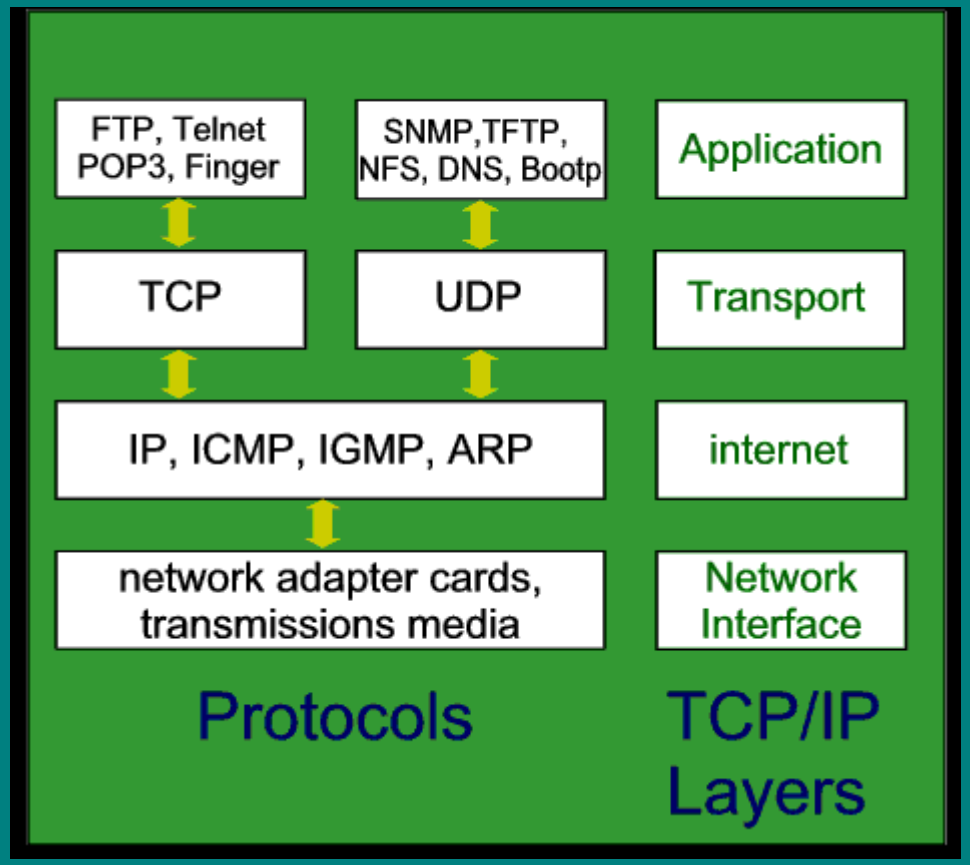
# OSI vs TCP/IP

- The OSI layer model is more complex than the TCP/IP model
- No commercially used system implements the OSI model
- Actually, no commercially used system ever implemented it
- The OSI model is usually used as a reference model for teaching and for academic purposes

# OSI vs TCP/IP: 7 vs 4 layers



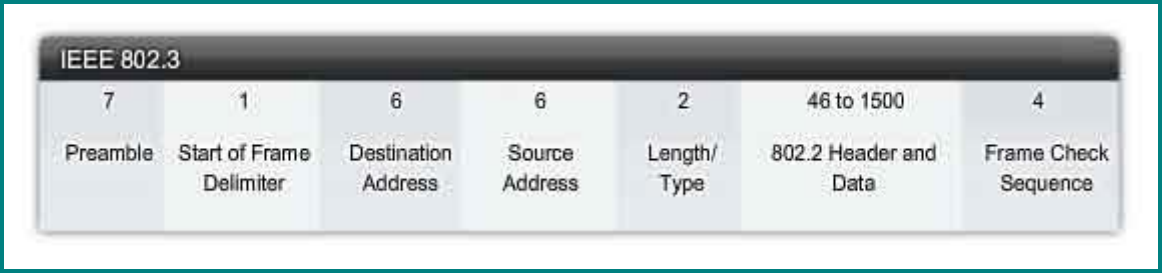
# The four TCP/IP layers



# Layer 2: Ethernet and MAC addresses

# The delivery system: Ethernet

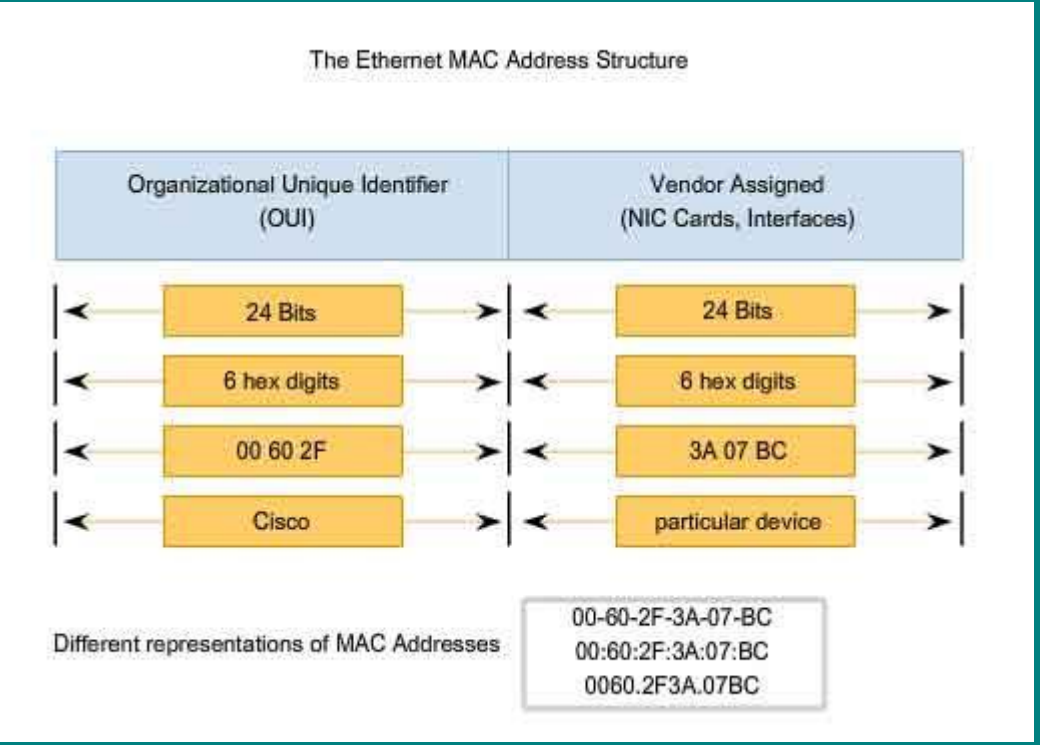
- Ethernet was first introduced in 1980 by the DIX consortium (“Digital Equipment Corporation/DEC, Intel and Xerox”) as an OPEN STANDARD and is defined in the IEEE standard 802.3 (IEEE is the “Institute of Electrical and Electronics Engineers”)
- Systems communicating over Ethernet divide a stream of data into individual packets called “frames”
- Ethernet frame size is up to 1518 bytes (without preamble and start of frame delimiter); IEEE 802.3ac extended the frame size to 1522 bytes in 1998:



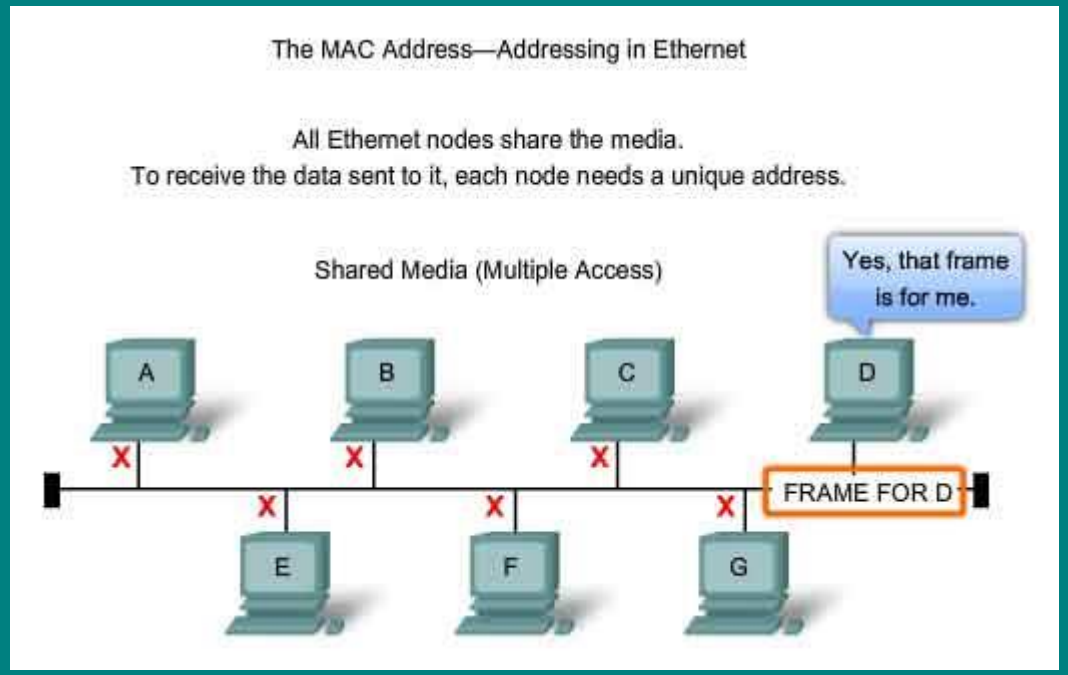
# MAC addresses

- MAC stands for “Media Access Control”; it has nothing to do with Apple Macintosh computers
- A MAC address has a length of 48 bits and is supposed to be unique – WORLD WIDE – and is used on Layer 2
- The standard (IEEE 802) format for printing MAC-48 addresses in human-friendly form is six groups of two hexadecimal digits, separated by hyphens (-) or colons (:), in transmission order (e.g. 01-23-45-67-89-ab or 01:23:45:67:89:ab)
- The first three octets identify the organization that issued the identifier and are known as the Organizationally Unique Identifier (OUI). The following three octets are freely assigned by that organization

# The 48-bit MAC address structure



# MAC addresses are used to identify hardware in the network



# Layer 3: IP addresses

# IP addresses

- An Internet Protocol address (IP address) is a numerical label assigned to each device (e.g., computer, printer) participating in a computer network that uses the Internet Protocol for communication
- An IP address serves two principal functions: host or network interface identification and location addressing. Its role has been characterized as follows:
- "A name indicates what we seek. An address indicates where it is. A route indicates how to get there”
- IP addresses are used on layer 3
- Currently, there are two versions of the IP protocol and IP addresses in use: IPv4 and IPv6

# Name, Address, Route

"A name indicates what we seek. An address indicates where it is. A route indicates how to get there”

```
C:\Users\wmaus>tracert 80.237.132.92

Tracing route to wp085.webpack.hosteurope.de [80.237.132.92]
over a maximum of 30 hops:

  1     3 ms     3 ms     4 ms  192.168.1.1
  2    50 ms    48 ms    55 ms  10.0.80.194
  3    83 ms    43 ms    46 ms  10.81.102.74
  4    35 ms    32 ms    46 ms  10.81.85.22
  5    75 ms    41 ms    22 ms  195.71.204.66
  6    63 ms    42 ms    39 ms  ae7-0.0002.corx.02.cgn.de.net.telefonica.de [62.53.8.234]
  7    55 ms    39 ms    39 ms  ae24-0.0002.corx.02.fra.de.net.telefonica.de [62.53.2.56]
  8    36 ms    39 ms    40 ms  bundle-ether2.0001.cord.01.off.de.net.telefonica.de [62.53.0.199]
  9    60 ms    40 ms    32 ms  bundle-ether1.0002.corp.01.off.de.net.telefonica.de [62.53.28.171]
 10    78 ms    53 ms    44 ms  et-1-1-2-u100.fra11-cr-polaris.bb.gdinf.net [80.81.192.239]
 11    79 ms    37 ms    41 ms  ae0.fra10-cr-antares.bb.gdinf.net [87.230.115.1]
 12    70 ms    38 ms    55 ms  ae2.cgn1-cr-nashira.bb.gdinf.net [87.230.114.4]
 13    35 ms    56 ms    48 ms  po250.sr-left.cgn1.dcnet-emea.godaddy.com [87.230.114.79]
 14     *        *        *     Request timed out.
 15    38 ms    35 ms    40 ms  wp085.webpack.hosteurope.de [80.237.132.92]

Trace complete.
```

# IPv4 addresses

- IPv4 uses 32 binary bits to create a single unique address on the network. An IPv4 address is expressed by four numbers separated by dots. Each number is the decimal representation for an eight-digit binary number, also called an octet
- Binary octets are not very user friendly, so instead of writing 11000000.10101000.00000000.00000001 we write 192.168.0.1
- IPv4 offers an address space of $2^{32}$; that's a bit more than 4.2 billion unique IP addresses, in the range from 0.0.0.0 to 255.255.255.255
- Apparently, that wasn't enough: We're running out of IPv4 addresses and that's why IPv6 was invented

# Reserved IP addresses

| | | |
|---|---|---|
| 0.0.0.0/8 | "This" Network | RFC 1122, Section 3.2.1.3 |
| 10.0.0.0/8 | Private-Use Networks | RFC 1918 |
| 127.0.0.0/8 | Loopback | RFC 1122, Section 3.2.1.3 |
| | „There's no place like 127.0.0.1" | |
| 169.254.0.0/16 | Link Local | RFC 3927 |
| | Automatic Private | |
| | IP Addressing (APIPA) | |
| 172.16.0.0/12 | Private-Use Networks | RFC 1918 |
| 192.0.0.0/24 | IETF Protocol Assignments | RFC 5736 |
| 192.0.2.0/24 | TEST-NET-1 | RFC 5737 |
| 192.88.99.0/24 | 6to4 Relay Anycast | RFC 3068 |
| 192.168.0.0/16 | Private-Use Networks | RFC 1918 |
| 198.18.0.0/15 | Network Interconnect | |
| | Device Benchmark Testing | RFC 2544 |
| 198.51.100.0/24 | TEST-NET-2 | RFC 5737 |
| 203.0.113.0/24 | TEST-NET-3 | RFC 5737 |
| 224.0.0.0/4 | Multicast | RFC 3171 |
| 240.0.0.0/4 | Reserved for Future Use | RFC 1112, Section 4 |
| 255.255.255.255/32 | Limited Broadcast | RFC 919, Section 7 |
| | | RFC 922, Section 7 |

# IPv6 addresses

- IPv6 uses 128 binary bits to create a unique address on the network. An IPv6 address is expressed by eight groups of hexadecimal numbers separated by colons, e.g. 2001:cdba:0000:0000:0000:0000:3257:9652
- Groups of numbers that contain all zeros are often omitted to save space, leaving a colon separator to mark the gap (as in 2001:cdba**::**3257:9652).
- Localhost (IPv4 127.0.0.1) e.g. now is ::1
- IPv6 offers an address space of $2^{128}$ (approximately $3.4\times10^{38}$) addresses; basically a number so large that I currently don't have the vocabulary for it
- An example ISP's IPv6 allocation is 2a02:588::/32
- That range is larger than the entire current Internet!

# We won‘t talk about IPv6 today

- IPv6 is NOT a topic of this session
- Many, many service providers, devices and software systems still don’t support IPv6
- For example, Windows Vista was the first Microsoft operating system that had IPv6 support – Windows XP (which still has an estimated market share of around 45%) and Windows Server 2003 do NOT support IPv6!
- The “spoken” de facto standard on the Internet still is IPv4
- The dilemma: Nobody uses IPv6 because nobody uses IPv6
- Did I say that the world still speaks IPv4?

# Why do we use IP addresses instead of MAC addresses?

- Killer argument: MAC addresses are not routable
- MAC addresses are an Ethernet feature; many devices don’t use or have MAC addresses
- MAC addresses are basically random when you receive a new piece of equipment and there’s only one MAC address per network interface
- IP addresses are structured and hierarchical and multiple IP addresses can be assigned to the same network interface
- The function of an IP address is similar to that of a telephone number which usually consists of a country code, an area code and a number for each phone

# Through the layers

# Ports

- Ports identify applications and services on a host
- When you see something like 192.168.0.1:80, the ":80" refers to a PORT on that specific IP address; in other words, an application (or service) is listening on that port for client requests
- The port number identifies a specific application/service that is running on the host
- The ports 0 to 1023 are called "system ports"; on operating systems like Unix/Linux/OS X, those system ports are also referred to as "privileged ports" because an application requires superuser privileges to use those ports
- Ports 1024 to 49151 are "user ports"
- 49152 to 65535 are "private ports"

# A few well known ports

- IANA (“Internet Assigned Numbers Authority”) maintains a list of “well known ports” (http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xml)
- Usually, a server listens on one these ports for client requests and then opens additional private ports for each client session
- Post Office Protocol 3 (POP3), which is used for fetching mails from an email server, listens on TCP port 110 for incoming requests
- SMTP (Simple Mail Transfer Protocol): TCP 25
- IMAP4 (Internet Message Access Protocol): TCP 143
- DNS (Domain Name Services): TCP AND UDP 53
- HTTP (HyperText Transfer Protocol): TCP 80
- FTP (File Transfer Protocol): TCP 21

# TCP and UDP

- Applications and Server services (Layer 4) use layer 3 TCP (Transmission Control Protocol) and/or UDP (User Datagram Protocol) ports to communicate over a network or to offer their server services to clients
- TCP offers error correction and a “delivery guarantee” with its flow control mechanisms; flow control determines when data needs to be re-sent, and stops the flow of data until previous packets are successfully transferred
- UDP has neither error correction nor flow control and offers speed instead, which is why media streaming services use UDP instead of TCP; by its nature, UDP is not reliable—messages may be lost or delivered out of order

# MAC, IP and ARP

- IP addresses are used on layer 3 to enable routable communication between hosts and networks
- There can be many IP addresses on a device but only one MAC address
- Ethernet (Layer 2) uses MAC addresses to communicate
- The Address Resolution Protocol (ARP) is a network protocol which maps a network layer protocol address (layer 3) to a data link layer (layer 2) hardware address
- When a host wants to send data to another host, ARP requests are sent to the network to resolve the MAC address behind the IP address
- Results are stored in the ARP cache

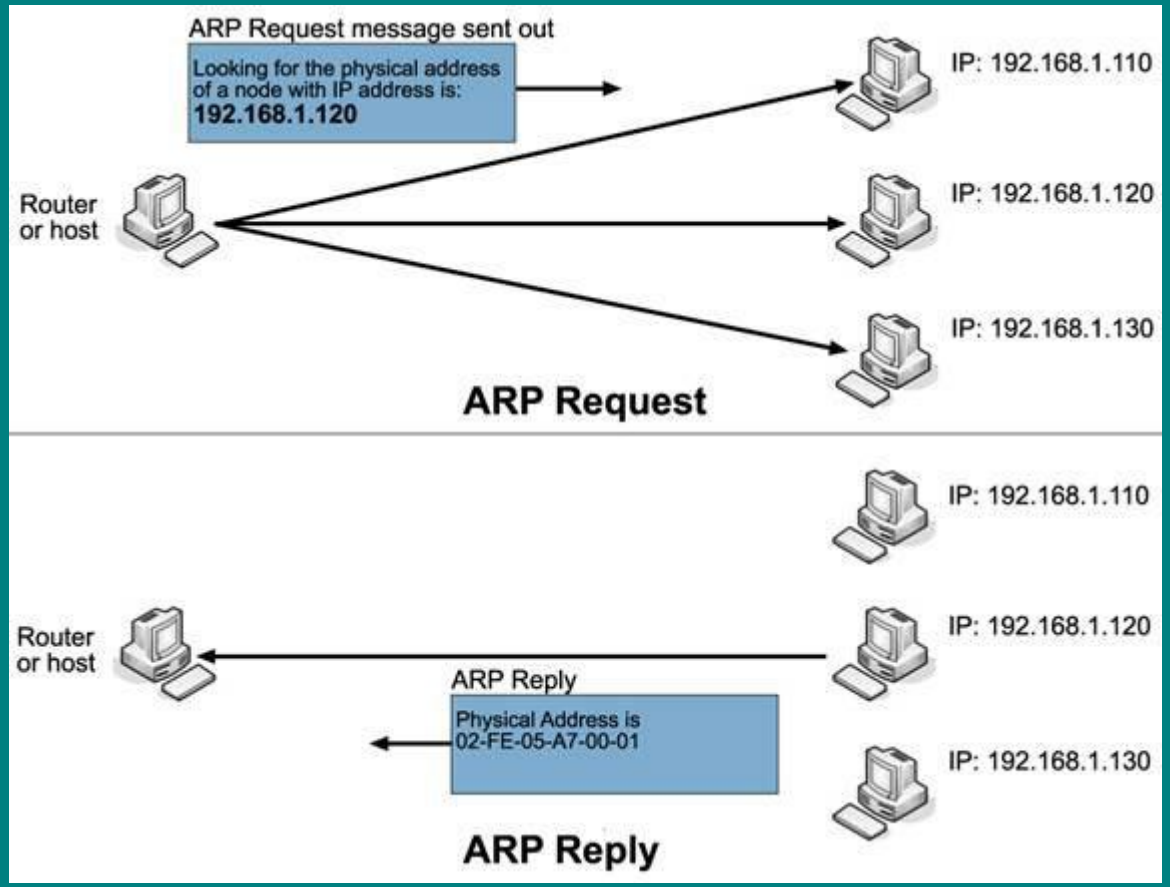
# ARP operation for local host

- When a network host sends data, a data packet containing the source IP address and the target IP address is passed down to the data link layer where it is taken and placed within a data link frame
- Based on the IP address (and the subnet mask), your computer should be able to figure out if the destination IP is a local IP or not

# ARP operation for local host

- If the IP is local, your computer will look in it's ARP table (a table where the responses to previous ARP requests are cached) to find the MAC address.
- If it's not there, then your computer will broadcast an ARP request to find out the MAC address for the destination IP. Since this request is broadcast, all machines on the LAN will receive it and examine the contents. If the IP address in the request is their own, they'll reply.
- On receiving this information, your computer will update it's ARP table to include the new information and will then send out the frame (addressed with the destination host's MAC address).Ethernet (Layer 2) uses MAC addresses to communicate

# ARP requests



# ARP operation for remote host

- If the IP is not local then the gateway (router) will see this (remember, the ARP request is broadcast so all hosts on the LAN will see the request)
- The router will look in it's routing table and if it has a route to the destination network, then it will reply with it's own MAC address
- This is only the case if your own computer doesn't know anything about the network topology. In most cases, your computer knows the subnet mask and has a default gateway set. Because of this, your own computer can figure out for itself that the packet is not destined for the local network

# ARP operation for remote host

- Instead, your computer will use the MAC address of the default gateway (which it will either have in it's ARP table or have to send out an ARP request for as outlined above)
- When the default gateway (router) receives the frame it will see that the MAC address matches it's own, so the frame must be for it
- The router will un-encapsulate the data link frame and pass the data part up to the network layer
- At the network layer, the router will see that the destination IP address (contained in the header of the IP packet) does not match it's own and will realize that this is a packet that is supposed to be routed

# ARP operation for remote host

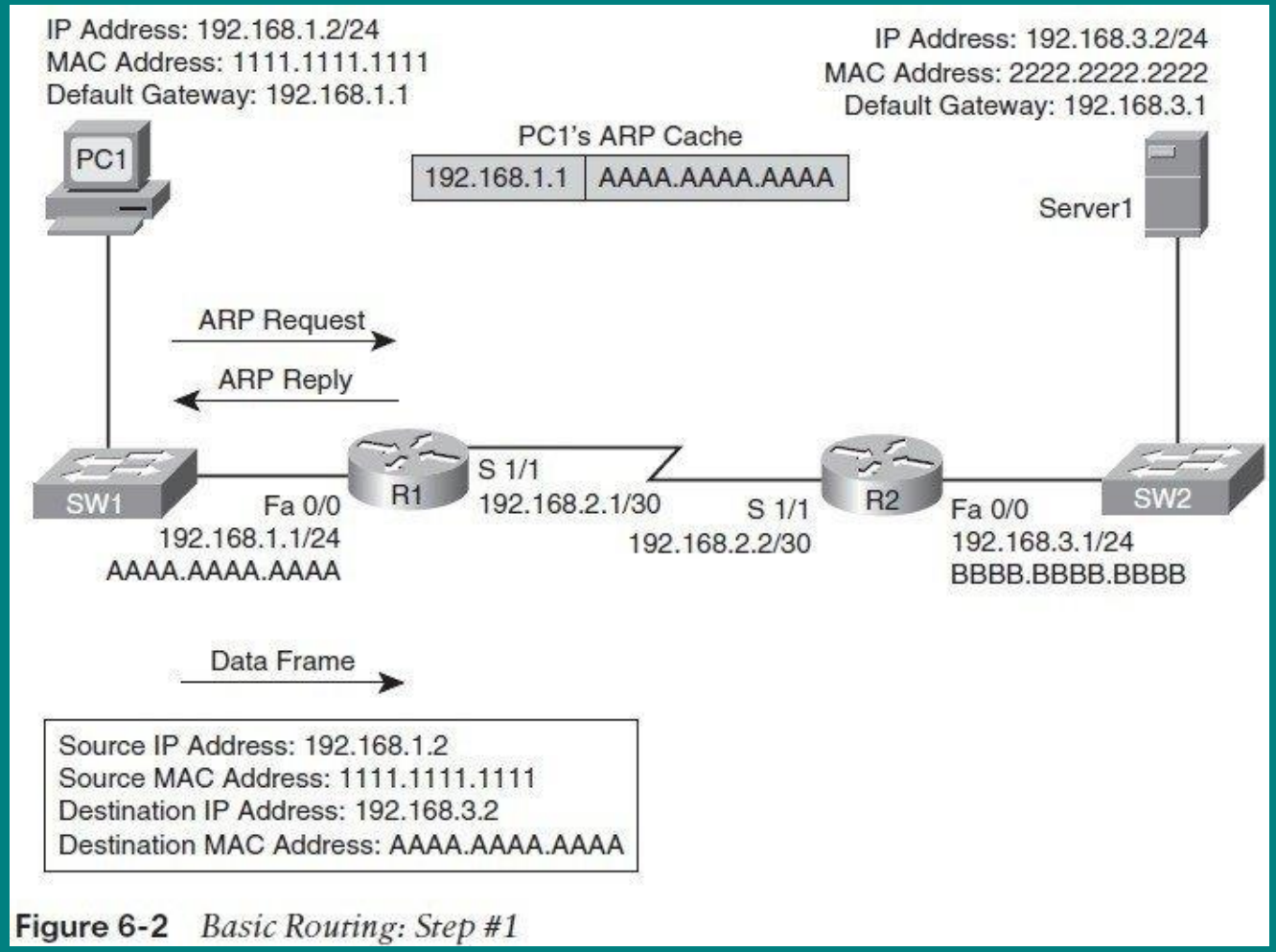



**Figure 6-2** *Basic Routing: Step #1*

# ARP operation for remote host

- The router will look in it's routing table for the closest match to the destination IP in order to figure out which interface to send the packet out on
- When a match is found, the router will create a new data link frame addressed to the next hop (and if the router doesn't know the hardware address for the next hop it will request it using the appropriate means for the technology in question)
- The data portion of this frame will contain the complete IP packet (where the destination IP address remains unchanged) and is sent out the appropriate interface

# ARP operation for remote host

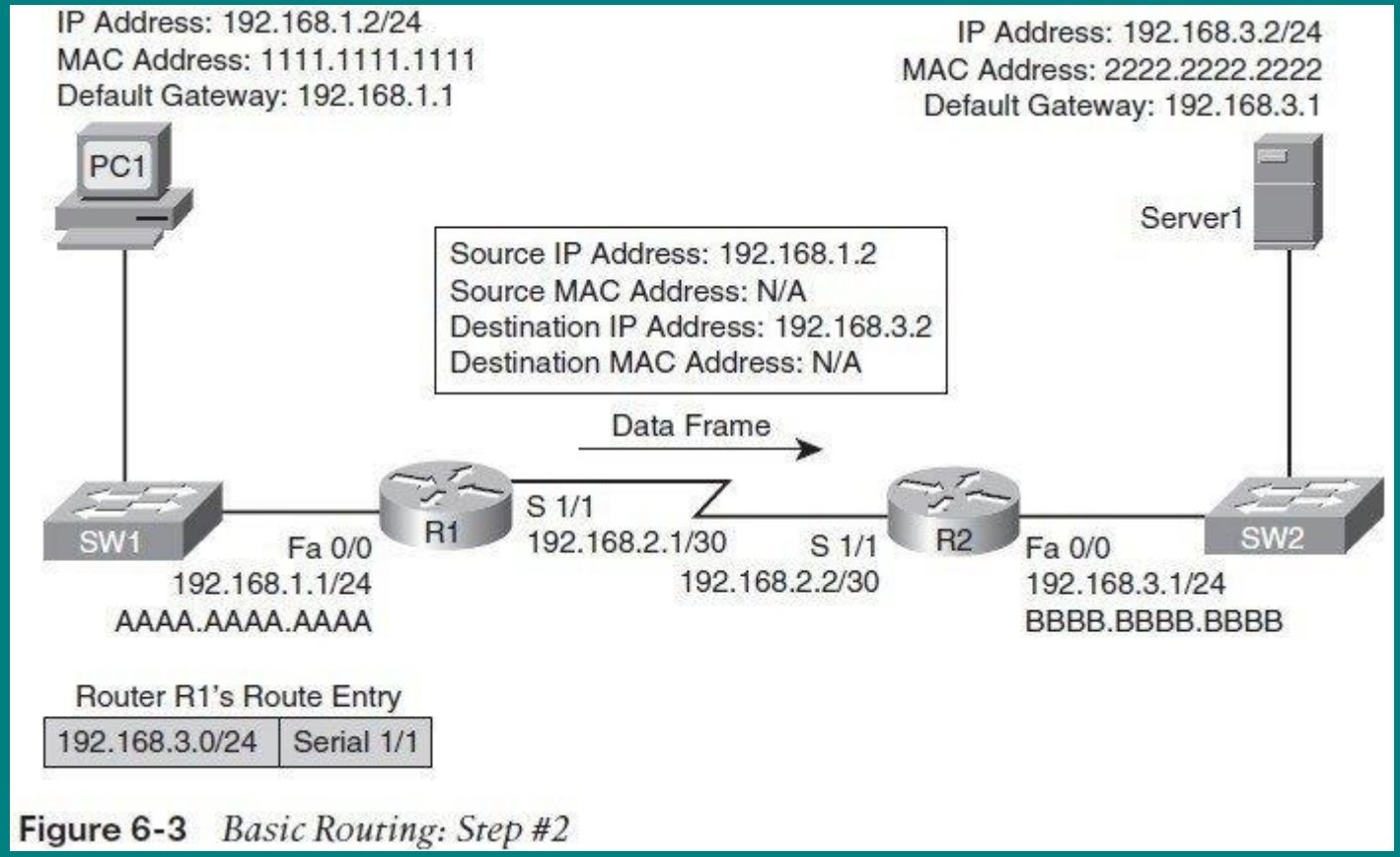



**Figure 6-3** *Basic Routing: Step #2*

# ARP operation for remote host

- This process will continue at each router along the way until the information reaches a router connected to the destination network
- This final router will see that the packet is addressed to a host that's on a directly connected network (the closest match you can get for an address, short of the packet being addressed to one self)
- It will send out an ARP request for MAC address of the destination IP (assuming it doesn't already have it in it's table) and then address the packet to the destination's MAC address

# ARP operation for remote host

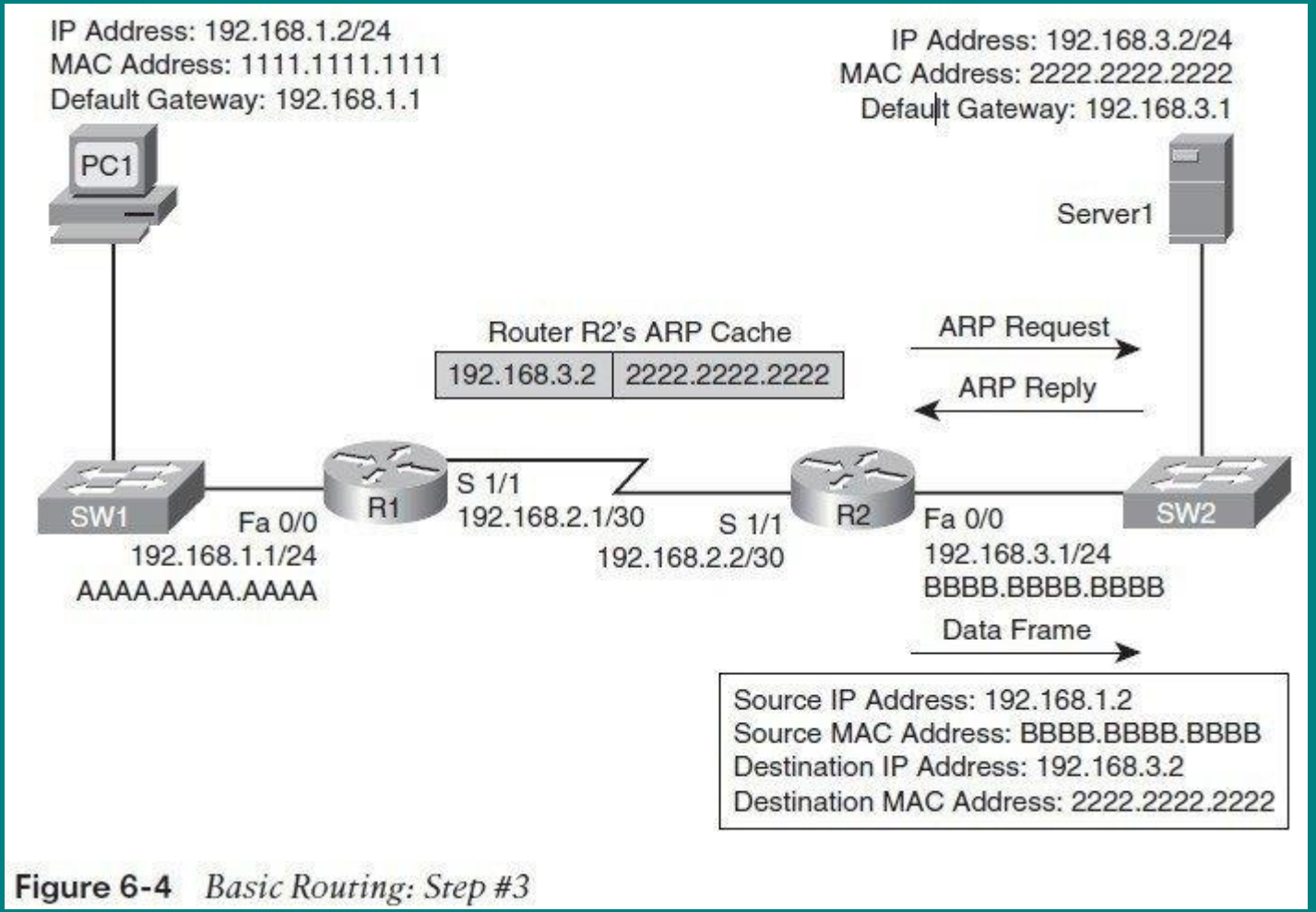



**Figure 6-4** *Basic Routing: Step #3*

# Broadcasts

- In computer networking, broadcasting refers to transmitting a packet that will be received by every device on the network
- In practice, the scope of the broadcast is limited to a broadcast domain
- A broadcast domain is a logical division of a computer network, in which all nodes can reach each other by broadcast at the data link layer

# Unicast and Multicast

- A Unicast is sending datagrams from one host to another single host identified by a unique IP address
- Multicast is the delivery of a message or information to a group of destination computers simultaneously in a single transmission from the source. Copies are automatically created in other network elements, such as routers, but only when the topology of the network requires it

# IP Multicast

- IP multicast is often employed in Internet Protocol (IP) applications of streaming media and Internet television
- IP multicast is a technique for one-to-many communication over an IP infrastructure in a network. It scales to a larger receiver population by not requiring prior knowledge of who or how many receivers there are
- Multicast uses network infrastructure efficiently by requiring the source to send a packet only once, even if it needs to be delivered to a large number of receivers. The nodes in the network take care of replicating the packet to reach multiple receivers only when necessary
- The most common transport layer protocol to use multicast addressing is User Datagram Protocol (UDP)

# VLANS

# Switches and Hubs

- Switches and Hubs are layer 2 devices that are being used to connect networking devices; they provide a logical bus over a physical star topology
- Hubs are “dumb”; they broadcast incoming packets to all of their ports
- Switches are more intelligent; they keep record of the MAC addresses behind their ports in so-called ARP-tables
- Switches use their ARP-table to determine where it makes sense to send the data packets
- Switches prevent packet collisions and improve the overall network performance and reliability

# Bridges

- A network bridge connects multiple network segments at the data link layer (Layer 2)
- A bridge and a switch are very much alike; a switch being a bridge with numerous ports
- Unlike routing, bridging makes no assumptions about where in a network a particular address is located. Instead, it depends on flooding and examination of source addresses in received packet headers to locate unknown devices. Once a device has been located, its location is recorded in a table where the source address is stored so as to avoid the need for further flooding
- The utility of bridging is limited by its dependence on flooding, and is thus only used in local area networks

# Virtual Local Area Netoworks (VLANs)

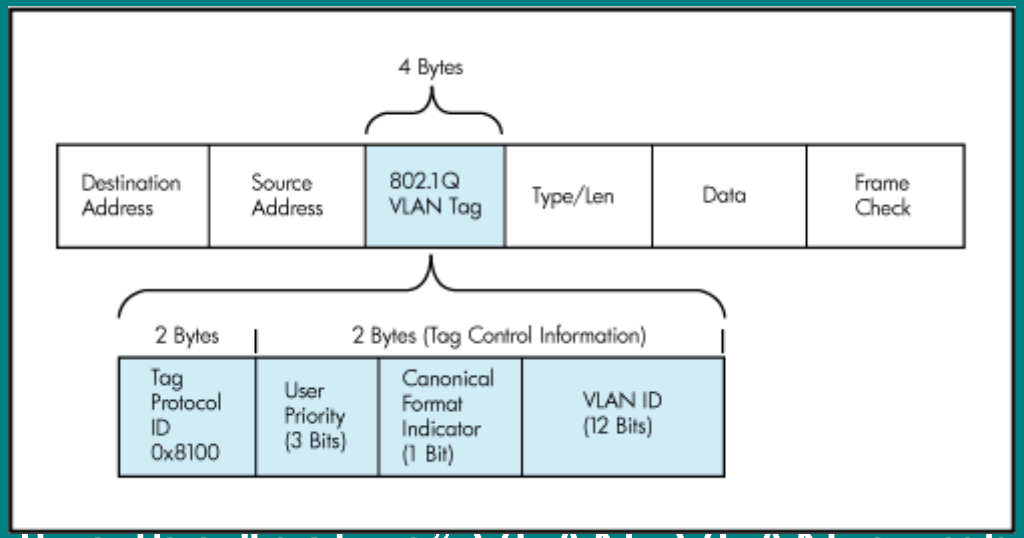
- A virtual local area network, virtual LAN or VLAN, is a group of hosts with a common set of requirements, which communicate as if they were attached to the same broadcast domain, regardless of their physical location
- A VLAN has the same attributes as a physical local area network (LAN), but it allows for end stations to be grouped together even if not on the same network switch
- VLAN membership can be configured through software instead of physically relocating devices or connections.
- To physically replicate the functions of a VLAN would require a separate, parallel collection of network cables and equipment separate from the primary network

# Virtual Local Area Netoworks (VLANs)

- VLANs are a layer 2 technology defined in the IEEE 802.1q standard – which explains the name of the command "`encapsulation dot1Q <vlanid>`" that you can read on some of Cisco routers ISPs use
- VLANs are configured on switch ports and in the switch's VLAN Trunk Protocol (VTP) database
- Trunks are used to pass VLAN information between switches
- Switches make sure that VLAN network traffic is separated from each other
- If you need to send packets from one VLAN to another, this can only be done via routing

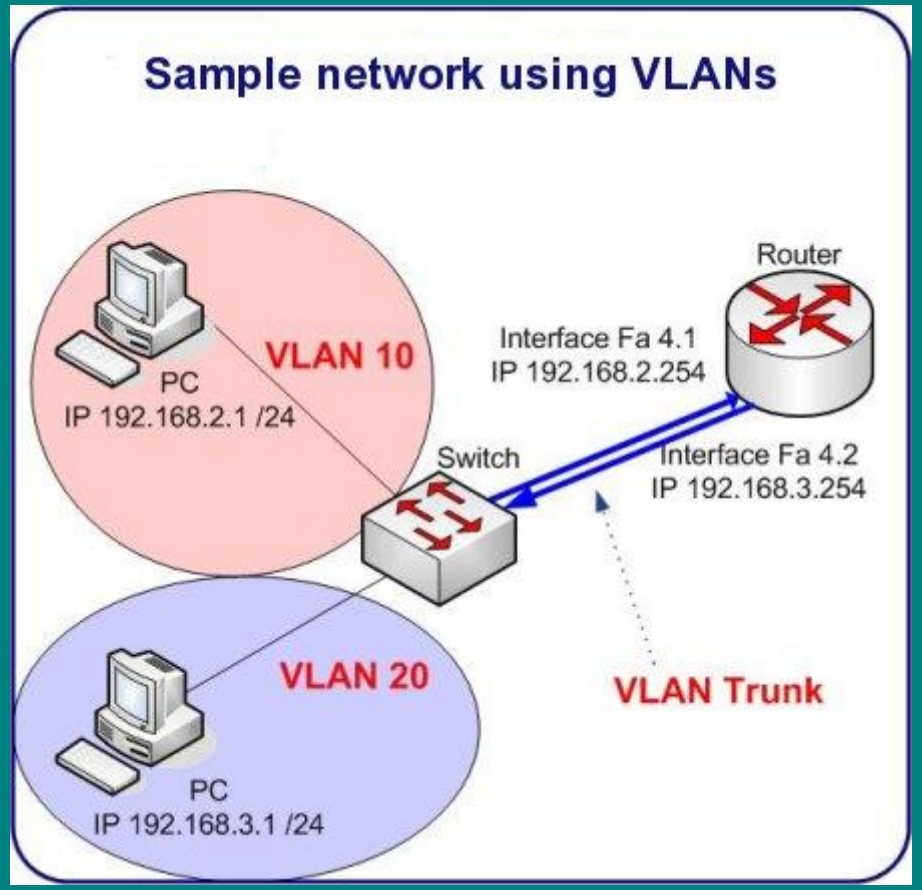
# VLAN Tagging

- When sent over a trunk port to another switch, a VLAN tag is inserted into the Ethernet frame that contains the VLAN id




- VLAN 1 by default is the "native" VLAN, VLAN numbers 2 to 1001 are free to use, Cisco internal VLANs begin at 1006. To avoid problems, stay within 2 to 1001

# A very simple VLAN setup



# Creating a VLAN on a Cisco switch

```
SW1>enable
SW1#conf t
SW1(config)#vlan 47
SW1(config-vlan)#name Test
SW1(config-vlan)#int fa0/1
SW1(config-if)#switchport access vlan 47
SW1(config-if)#switchport mode access
SW1(config-if)#end
SW1#write memory
```

# Classful and classless networking

# Classful networking

- Originally, IPv4 addresses were grouped into three classes A, B and C, which is called “classful networking”
- A class A network has 16,277,216 possible addresses (addresses, not hosts) per network, the range is from 1.0.0.0 to 126.255.255.255
- A class B network has 65,536 possible addresses per network in the IP address range from 128.0.0.0 to 191.255.255.255
- A class C network has 256 possible addresses per network in the address range from 192.0.0.0 to 223.255.255.255

# Classful networking examples

- A host with the IP address 10.47.11.1 is in the class A network 10.0.0.0
- A host with the IP address 172.16.20.1 is in the class B network 172.16.0.0
- A host with the IP address 192.168.128.95 is in the class C network 192.168.128.0

# Classful networking

- When the Internet grew bigger, it became quickly obvious that classful networking wasted IP addresses
- For example, the Massachussettes Institute of Technology (MIT) owns an entire Class A network – now what does a university need 16 million IP addresses for? However, at the time class A networks were easier to get than it is today; more info on MIT's network range can be found here: http://libstaff.mit.edu/colserv/digital/ordering/ip.html
- Cisco routers still support classful networking, and in the case of certain routing protocols like RIP or EIGRP even use classful networking as their default setting
- Classful networking is useful for creating summary addresses (e.g. you might want to send all traffic to private networks to one specific router)

# Classless networking

- Classless Inter-Domain Routing (CIDR) is a method for allocating IP addresses and routing Internet Protocol packets
- The Internet Engineering Task Force introduced CIDR in 1993 to replace the previous addressing architecture of classful network design in the Internet
- Their goal was to slow the growth of routing tables on routers across the Internet, and to help slow the rapid exhaustion of IPv4 addresses.

# Variable Length Subnet Mask (VLSM)

- Classless Inter-Domain Routing is based on variable-length subnet masking (VLSM), which allows a network to be divided into variously-sized subnets, providing the opportunity to size a network more appropriately for local needs
- Variable-length subnet masks are mentioned in RFC 950

# Variable Length Subnet Mask (VLSM)

- IPv4 CIDR blocks are identified using a syntax similar to that of IPv4 addresses: a dotted-decimal address, followed by a slash, then a number from 0 to 32, e.g., a.b.c.d/n
- The dotted decimal portion is the IPv4 address. The number following the slash is the prefix length, the number of shared initial bits, counting from the most-significant bit of the address
- When emphasizing only the size of a network, the address portion of the notation is usually omitted. Thus, a /20 is a CIDR block with an unspecified 20-bit prefix

# Variable Length Subnet Mask (VLSM)

- A class A network in CIDR notation is a /8
- A class B network in CIDR notation is a /16
- A class C network in CIDR notation is a /24
- Routes smaller than /24 are not advertised on the Internet
- The smallest network allocation for public IPv4 addresses that RIPE grants is /21 (more than 2000 IP addresses or 8 class C networks) (ftp://ftp.bme.hu/documents/ripe/ripe-328.pdf)

# Subnet masks

- A subnet mask is a bitmask that encodes the prefix length in quad-dotted notation: 32 bits, starting with a number of 1 bits equal to the prefix length, ending with 0 bits, and encoded in four-part dotted-decimal format
- A subnet mask encodes the same information as a prefix length, but predates the advent of CIDR
- Examples
- /24 in CIDR notation equals to the subnet mask 255.255.255.0
- /30 in CIDR equals to 255.255.255.252
- /28 in CIDR equals to 255.255.255.240

# IPv4 CIDR Chart

| Netmask | Netmask (binary) | CIDR | Notes |
|---|---|---|---|
| 255.255.255.255 | 11111111.11111111.11111111.11111111 | /32 | Host (single addr) |
| 255.255.255.254 | 11111111.11111111.11111111.11111110 | /31 | Unuseable |
| 255.255.255.252 | 11111111.11111111.11111111.11111100 | /30 | 2 useable – the typical „transfer network“ size |
| 255.255.255.248 | 11111111.11111111.11111111.11111000 | /29 | 6 useable |
| 255.255.255.240 | 11111111.11111111.11111111.11110000 | /28 | 14 useable |
| 255.255.255.224 | 11111111.11111111.11111111.11100000 | /27 | 30 useable |
| 255.255.255.192 | 11111111.11111111.11111111.11000000 | /26 | 62 useable |
| 255.255.255.128 | 11111111.11111111.11111111.10000000 | /25 | 126 useable |
| 255.255.255.0 | 11111111.11111111.11111111.00000000 | /24 | "Class C" 254 useable |
| 255.255.254.0 | 11111111.11111111.11111110.00000000 | /23 | 2 Class C's |
| 255.255.252.0 | 11111111.11111111.11111100.00000000 | /22 | 4 Class C's |
| 255.255.248.0 | 11111111.11111111.11111000.00000000 | /21 | 8 Class C's |
| 255.255.240.0 | 11111111.11111111.11110000.00000000 | /20 | 16 Class C's |
| 255.255.224.0 | 11111111.11111111.11100000.00000000 | /19 | 32 Class C's |
| 255.255.192.0 | 11111111.11111111.11000000.00000000 | /18 | 64 Class C's |
| 255.255.128.0 | 11111111.11111111.10000000.00000000 | /17 | 128 Class C's |
| 255.255.0.0 | 11111111.11111111.00000000.00000000 | /16 | "Class B„ 65534 useable |
| 255.254.0.0 | 11111111.11111110.00000000.00000000 | /15 | 2 Class B's |
| 255.252.0.0 | 11111111.11111100.00000000.00000000 | /14 | 4 Class B's |
| 255.248.0.0 | 11111111.11111000.00000000.00000000 | /13 | 8 Class B's |
| 255.240.0.0 | 11111111.11110000.00000000.00000000 | /12 | 16 Class B's |
| 255.224.0.0 | 11111111.11100000.00000000.00000000 | /11 | 32 Class B's |
| 255.192.0.0 | 11111111.11000000.00000000.00000000 | /10 | 64 Class B's |
| 255.128.0.0 | 11111111.10000000.00000000.00000000 | /9 | 128 Class B's |
| 255.0.0.0 | 11111111.00000000.00000000.00000000 | /8 | "Class A„ 16777214 useable |
| 254.0.0.0 | 11111110.00000000.00000000.00000000 | /7 | |
| 252.0.0.0 | 11111100.00000000.00000000.00000000 | /6 | |
| 248.0.0.0 | 11111000.00000000.00000000.00000000 | /5 | |
| 240.0.0.0 | 11110000.00000000.00000000.00000000 | /4 | |
| 224.0.0.0 | 11100000.00000000.00000000.00000000 | /3 | |
| 192.0.0.0 | 11000000.00000000.00000000.00000000 | /2 | |
| 128.0.0.0 | 10000000.00000000.00000000.00000000 | /1 | |
| 0.0.0.0 | 00000000.00000000.00000000.00000000 | /0 | IP space |

# Determining the network prefix

- An IPv4 network mask consists of 32 bits, a sequence of ones (1) followed by a block of 0s. The trailing block of zeros (0) designates that part as being the host identifier
- The following example shows the separation of the network prefix and the host identifier from an address (192.168.5.130) and its associated /24 network mask (255.255.255.0)
- The binary AND operator is used to determine the network prefix

| | Binary form | Dot-decimal notation |
|---|---|---|
| IP address | 11000000.10101000.00000101.10000010 | 192.168.5.130 |
| Subnet mask | 11111111.11111111.11111111.00000000 | 255.255.255.0 |
| Network prefix | 11000000.10101000.00000101.00000000 | 192.168.5.0 |
| Host part | 00000000.00000000.00000000.10000010 | 0.0.0.130 |

# The binary AND operator

- A bitwise AND takes two binary representations of equal length and performs the logical AND operation on each pair of corresponding bits
- The result in each position is 1 if the first bit is 1 and the second bit is 1; otherwise, the result is 0
- ```
  0101 (decimal 5)
  ```
- ```
  AND 0011 (decimal 3)
  ```
- ```
  = 0001 (decimal 1)
  ```
  ```

# When is the default gateway used?

- How does the host know whether an IP address is local or remote and how does it know if a data packet must be routed to a remote network?
- Answer: It uses the bitwise AND operator and its own subnet mask on the target IP address to determine a network prefix for the target IP address;
- If the network prefix of the target IP address matches the host's own network prefix, then the target IP address must be local
- If not, the packet must be routed and accordingly will be sent to the default gateway

# Example I

- Source IP: 192.168.0.112/24
- Subnetmask: 255.255.255.0

- Source IP: 11000000 10101000 00000000 01110000
- Subnet Mask: 11111111 11111111 11111111 00000000
- Source-Prefix: 11000000 10101000 00000000 00000000

- Target IP: 192.168.0.47

- Target IP: 11000000 10101000 00000000 00101111
- Subnet Mask: 11111111 11111111 11111111 00000000
- Target-Prefix: 11000000 10101000 00000000 00000000

  - In this case, the Source-Prefix equals the Target-Prefix, so it must be a local address

# Example II

- Source IP: 192.168.3.112/24
- Subnetmask: 255.255.255.0

- Source IP: 11000000 10101000 00000011 01110000
- Subnet Mask: 11111111 11111111 11111111 00000000
- Source-Prefix: 11000000 10101000 00000011 00000000

- Target IP: 192.168.0.47

- Target IP: 11000000 10101000 00000000 00101111
- Subnet Mask: 11111111 11111111 11111111 00000000
- Target-Prefix: 11000000 10101000 00000000 00000000

  - In this case, the Source-Prefix does NOT equal the Target-Prefix, so it must be a remote address

# Routing

# Autonomous Systems

- Within the Internet, an Autonomous System (AS) is a collection of connected Internet Protocol routing prefixes under the control of one or more network operators that presents a common, clearly defined routing policy to the Internet
- Internet Service Providers (ISP) must have an officially registered Autonomous System Number (ASN) assigned by IANA
- The example ISP’s ASN is 39151
- The example ISP’s public subnet prefixes are:
- 87.238.112.0/21 (DE-EXISP-20051222, >2000 IP addresses)
- 91.151.144.0/20 (DE-EXISP-20061213, >4000 IP addresses)

# Routers

- Routers pass data between multiple networks
- Routers are layer 3 devices (there are also layer 3 switches on the market today)
- Routers are essentially computers optimized for handling data packets
- They attempt to send packets from the source to the target in the fastest way possible (and that is not always the shortest path)
- When a router receives a packet destined for a point outside of its own networks, it examines its routing table to find a suitable route to the destination network; if it does not have such a route the router might still have a default gateway of its own to which it will send the packet

# Routes and routing protocols

- Routers can have static routes to various destinations
- They can have a default route that they will use for all destinations for which they do not have explicit routing information
- Routers will always use the most specific route to the destination that they find in their routing table
- Routers use various routing protocol through which they can learn routes from other routers in their neighborhood
- Interior Gateway Protocols (IGP) to be used within an Autonomous System (AS) would be RIP, EIGRP (Cisco proprietary), OSPF
- IS-IS and BGP would be used between AS

# Routes and routing protocols

- When a Cisco router learns the same route through different routing protocols, it will put the route with the smallest administrative distance in its routing table
- The default administrative distance values in Cisco iOS are:

| | |
|---|---|
| Connected | 0 |
| Static | 1 |
| eBGP | 20 |
| EIGRP (internal) | 90 |
| IGRP | 100 |
| OSPF | 110 |
| IS-IS | 115 |
| RIP | 120 |
| EIGRP (external) | 170 |
| iBGP | 200 |
| EIGRP summary route | 5 |

# Static routes, OSPF & BGP

- In a heterogenous network infrastructure, vendor-specific and proprietary protocols like Cisco's EIGRP cannot be used; e.g. Mikrotik routers don't talk EIGRP
- BGP is –THE– routing protocol used on the Internet and all service providers "speak" it with each other so naturally we support it as well
- Unlike OSPF, EIGRP or RIP, BGP is a pure TCP protocol and thus requires full IP connectivity between the participating ("neighboring") routers in order to function
- In other words, BGP requires that either another routing protocol like OSPF already runs on the network or that the connectivity has been established via static routes

# Static routes

# Static route configuration on Cisco

- `ip route 41.191.118.0 255.255.255.0 GigabitEthernet0/2.60 91.151.145.244 name GEOLINK_iDirect_2`
- `ip route 41.191.118.0 255.255.255.0 Null0 254 name GEOLINK_BGP_announcement-20110518`

  - This first statement says that the target network 41.191.118.0/24 is available over the router's virtual interface Ge0/2.60 which in turn will send the traffic destined to this network to the router 91.151.145.244

# Static route configuration on Cisco

- `ip route 41.191.118.0 255.255.255.0 GigabitEthernet0/2.60 91.151.145.244 name GEOLINK_iDirect_2`
- `ip route 41.191.118.0 255.255.255.0 Null0 254 name GEOLINK_BGP_announcement-20110518`

  - The second statement configures a backup route to the same network with a metric of 254 (which is very bad) and all traffic would be sent to the null device, the router's digital garbage bin
  - The second statement is important for BGP: Per default, BGP would drop the network 41.191.118.0 from the routing table when the interface Ge0/2.60 shuts down or when the peer router at 91.151.145.244 becomes unavailable; in this case, the network 41.191.118.0 would no longer be advertised and disappear from the Internet

# OSPF

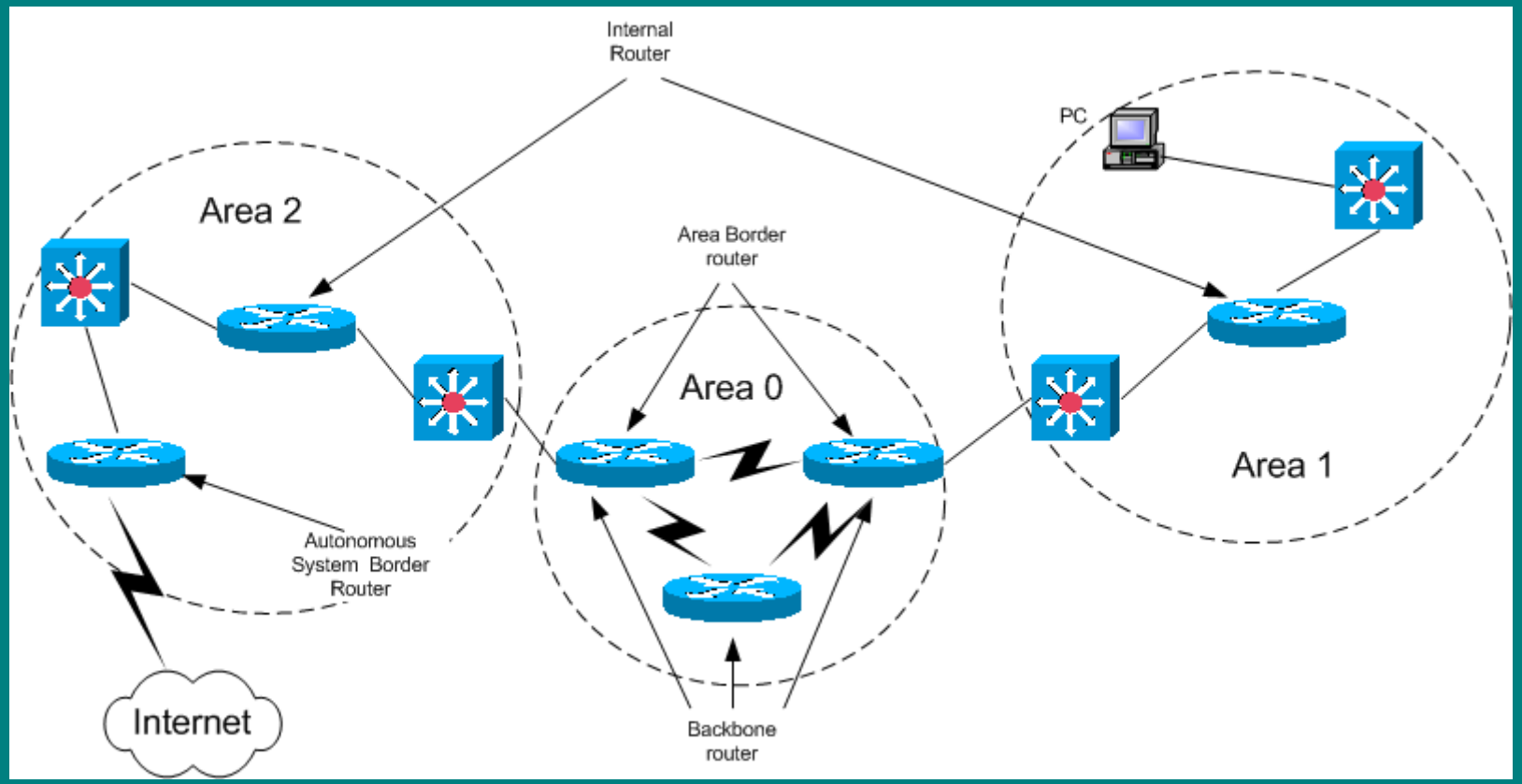
# OSPF

- OSPF stands for “Open Shortest Path First” and is a so-called “adaptive link state routing protocol”
- It is an Interior Gateway Protocol (IGP) designed to be used within a single Autonomous System (AS)
- It is the most widely used IGP in today’s enterprise environments
- As the name implies, OSPF is an OPEN standard and it determines the best route to the destination by applying Dijkstra’s “Shortest Path First” algorithm to the router’s link state database (LSDB)
- The LSDB collects the information contained in Link State Advertisements (LSA), which are always sent when the topology of the network changes

# OSPF areas

- OSPF uses a concept of routing “areas”, where area 0 is the designated backbone area
- Area 0, the backbone area, is supposed to be a pure transit area and its function is to quickly and efficiently move IP packets
- All other areas are deemed “normal areas” and their primary function is to connect users and resources
- According to Cisco, there shouldn’t be more than 50 routers in one OSPF area
- The example ISP currently uses around a dozen routers in the backbone area, so we are far away from that theoretical limit and definitely only need one OSPF area

# Example for OSPF areas



# OSPF router roles

- In an OSPF structure, routers play different roles as for example “autonomous system boundary router”, “area border router”, “internal router” and “backbone router”
- “Autonomous System Boundary Routers” (ASBR) is connected to more than one routing protocol and exchanges routing information with routers in other protocols
- “Area Border Routers” (ABR) connect one or more network areas to the backbone area
- “Backbone Routers” are all routers connected to the backbone
- “Internal routers” are routers that have all their interfaces in one OSPF area

# OSPF Designated Router

- “Designated Routers” (DR) are elected router interfaces and exist for the purpose of reducing network traffic by providing a source for routing updates
- The DR maintains a complete topology table of the network and sends the updates to the other routers
- All routers in a multi-access network segment will form a slave/master relationship with the DR. They will form adjacencies with the DR and BDR only. Every time a router sends an update, it sends it to the DR and BDR
- The DR will then send the update out to all other routers in the area
- This way all the routers do not have to constantly update each other, and can rather get all their updates from a single source

# OSPF and the Dijkstra calculation

- Edsger Dijkstra's SPF algorithm assigns "costs" (also known as metric) to each link in the network and then sums the cost to each destination from the router
- The path to the destination with the lowest cost is put in the routing table; costs are re-calculated with every link state change
- OSPF bases the cost on the configured bandwidth of the interfaces (which can be overridden manually)
- Default costs in OSPF:
- 100 Mbit/s FastEthernet: 1
- 10 Mbit/s Ethernet: 10
- E1 line (2 Mbit/s): 48
- 56 kb/s serial link: 1785

# OSPF in an ISP network

- In its current implementation, an ISP uses two dedicated networks for OSPF traffic (e.g. VLAN 91 and VLAN 92) but only one OSPF area: Area 0
- All routers are members of OSPF area 0 and use only one OSPF process with the process id 1
- By design, OSPF is supposed to serve only ONE purpose in our network: To make the loopback, infrastructure and BGP peering neighbor addresses of  our routers known and accessible on the network
- Those addresses are required for BGP to establish peering neighborhoods between the various routers and systems (OSPF allows them to talk to each other)
- In an ISP network environment, OSPF should not be used beyond that scope

# OSPF in an ISP network

- An ISP's backbone routers have OSPF priorities configured to make sure that Backbone Routers 1 and 2 will always "win" the DR election process with such  configurations on their OSPF interfaces:

```
interface GigabitEthernet0/1.91
 description BBLAN #1
 encapsulation dot1Q 91
 ip address 91.151.144.33 255.255.255.240
 ip ospf priority 128
```

- In the case of OSPF, the highest priority value wins the DR election

# OSPF in an ISP network

- An example ISP uses the network 91.151.144.32/28 on VLAN 91 as the first OSPF network
- The second OSPF network is 91.151.144.48/28 on VLAN 92
- You only need one network to make OSPF work, but this setup provides redundancy in the case of failure of one network
- The “network” command in Cisco’s OSPF configuration section does not only advertise a network, but also enables ALL interfaces that have an address in the specified range to use the OSPF protocol:

```
router ospf 1
 router-id 91.151.144.1
 network 91.151.144.32 0.0.0.15 area 0
 network 91.151.144.48 0.0.0.15 area 0
```

# OSPF wild card bits

- The “network” command uses “wild card bits” to match what IP addresses belong to the advertised network range
- A 0 (zero) in a wildcard mask means to check the corresponding bit in the address for an exact match
- A 1 (one) in a wildcard mask means to ignore the corresponding bit in the address—can be either 1 or 0
- An octet of all zeros means that the octet has to match the address exactly. An octet of all ones means that the octet can be ignored

```
router ospf 1
 router-id 91.151.144.1
 network 91.151.144.32 0.0.0.15 area 0
 network 91.151.144.48 0.0.0.15 area 0
```

# OSPF wild card bits

- An example for the use of wild card bits:
- 172.16.8.0  0.0.7.255

```
172.168.8.0 =       10101100.00010000.00001000.00000000
                      |-MUST MATCH EXACTLY--|
0.0.0.7.255 =       00000000.00000000.00000111.11111111
result =            10101100.00010000.00001xxx.xxxxxxxx

                                      00001000
                                      00001001
                                      00001010
                                      00001011
                                      00001100
                                      00001101
                                      00001111
00001xxx =          00001000 to 00001111 = 8 - 15
xxxxxxxx =          00000000 to 11111111 = 0 - 255
```

- Anything between 172.16.8.0 and 172.16.15.255 will match the example statement

# OSPF configuration on Cisco á la maison

```
interface Loopback0
 description Management Loopback
 ip address 91.151.144.5 255.255.255.255

interface FastEthernet0/0.91
 description OSPF-Lan 1
 encapsulation dot1Q 91
 ip address 91.151.144.37 255.255.255.240
 ip ospf authentication message-digest
 ip ospf message-digest-key 1 md5 7 08254D5D000A111317190F013E2E282720
 ip ospf cost 100
 ip ospf hello-interval 3
 ip ospf dead-interval 10
 ip ospf priority 32

 interface FastEthernet0/1.92
 description OSPF-Lan 2
 encapsulation dot1Q 92
 ip address 91.151.144.53 255.255.255.240
 ip ospf authentication message-digest
 ip ospf message-digest-key 1 md5 7 110D18161E011F08013828213C36392D00
 ip ospf cost 100
 ip ospf hello-interval 3
 ip ospf dead-interval 10
 ip ospf priority 32

router ospf 1
 router-id 91.151.144.5
 log-adjacency-changes
 passive-interface Loopback0
 network 91.151.144.5 0.0.0.0 area 0
 network 91.151.144.32 0.0.0.15 area 0
 network 91.151.144.48 0.0.0.15 area 0
```

# OSPF in an ISP network

- There is MUCH more to OSPF than we have covered here
- OSPF also has many more possibilities and features than this example ISP's uses
- And OSPF certainly has more features than any ISP needs
- Remember:
- OSPF is designed to be used WITHIN an autonomous system
- OSPF is only needed in an ISP network as an internal infrastructure foundation for monitoring and BGP
- The mission critical routing protocol in an ISP's network is BGP

# BGP

# BGP

- BGP stands for “Border Gateway Protocol”
- It is a so-called path vector protocol and is being used to connect Autonomous Systems with each other
- BGP does not use traditional Interior Gateway Protocol (IGP) metrics, but makes routing decisions based on path, network policies and/or rule-sets. For this reason, it is more appropriately termed a reach-ability protocol rather than routing protocol; unless configured otherwise, the path through the lowest number of Anonymous Systems is preferred
- BGP maintains a table of IP networks or 'prefixes' which designate network reach-ability among autonomous systems (AS)
- It is the de-facto standard for routing on the Internet

# BGP

- It is “the Linux under the routing protocols”: It is extremely customizable and everything must be configured manually
- Again: There is not one automatism in BGP and all values that are used to calculate the best path to a destination can be manipulated
- BGP is known to adjust only slowly to network changes
- Establishing neighborhood relationships with peer routers might also be rather slow
- It is robust, reliable and guaranteed to be free of routing loops

# BGP uses TCP

- BGP neighbors, called peers, are established by manual configuration between routers to create a TCP session on port 179
- A BGP speaker will periodically (every 30 seconds) send 19-byte keep-alive messages to maintain the connection.
- Among routing protocols, BGP is unique in using TCP as its transport protocol – and it is the reason why BGP usually relies on the existence of another internal routing protocol on its Autonomous System (AS) to ensure functional TCP connectivity between the peers

# IBGP and EBGP

- When BGP runs between two peers in the same autonomous system (AS), it is referred to as Internal BGP (IBGP or Interior Border Gateway Protocol)
- When it runs between autonomous systems, it is called External BGP (EBGP or Exterior Border Gateway Protocol)
- Routers on the boundary of one AS exchanging information with another AS are called border or edge routers
- In the Cisco operating system, IBGP routes have an administrative distance of 200 and that of EBGP is 20; IBGP is thus less preferred than either external BGP or any interior routing protocol

# AS numbers and process ids

- Unlike OSPF, in BGP the process id plays a more important role than just assigning a number to the process so that it can be identified by the operating system
- In BGP, the process is used to identify the Autonomous System to which the router belongs
- The router's BGP process must be configured with either an official or a private Autonomous System Number (ASN)
- AS numbers range from 1 to 65535, the range from 64512 to 65535 is reserved for private AS
- The example ISP's registered ASN is 39151

# A very basic IBGP configuration

- ROUTER 1 on IP address 10.1.1.1:

- Router1(config)# router bgp 39151
- Router1(config-router)# neighbor 10.1.1.2 remote-as 39151
- Router1(config-router)# neighbor 10.1.1.2 update-source Loopback 0
- Router1(config-router)# network 192.168.0.0 mask 255.255.255.0

- ROUTER 2 on IP address 10.1.1.2:

- Router2(config)# router bgp 39151
- Router2(config-router)# neighbor 10.1.1.1 remote-as 39151
- Router2(config-router)# neighbor 10.1.1.1 update-source Loopback 0
- Router2(config-router)# network 192.168.2.0 mask 255.255.255.0

- Using a loopback address as the update source must not be configured but it adds resiliency to IBGP sessions because the loopback interface will always be available while physical interfaces might be shut down or otherwise fail;
- unavailable interfaces will cause a change of the routing table;
- also, when a BGP update package comes from an IP address that the peer has not established a neighborhood with, that package is simply discarded;
- both parties must know their peer‘s IP addresses
- Remember: OSPF makes sure that loopback addresses are reachable

# A very basic EBGP configuration

- **ROUTER 1 on IP address 91.151.144.1:**

- `Router1(config)# router bgp 39151`
- `Router1(config-router)# neighbor 87.238.112.1 remote-as 65020`

- *(A password is not required, but adds some additional safety; the neighborhood is only established when both peers use the same password.)*
- `Router1(config-router)# neighbor 87.238.112.1 password verysecretpassword`

- *(next-hop-self forces BGP to use ist own IP address as the next hop address for each network that Router 1 advertises to its neighbor )*
- `Router1(config-router)# neighbor 87.238.112.1 next-hop-self`

- *(Advertise the network 91.151.155.0 with the subnet mask 255.255.255.0; those are NOT wild cards but real subnet masks)*
- `Router1(config-router)# network 91.151.155.0 mask 255.255.255.0`

- **ROUTER 2 on IP address 87.238.112.1:**

- `Router2(config)# router bgp 65020`
- `Router2(config-router)# neighbor 91.151.144.1 remote-as 39151`
- `Router2(config-router)# neighbor 91.151.144.1 password verysecretpassword`
- `Router2(config-router)# network 87.238.113.0 mask 255.255.255.0`

# Manipulating path calculations with route maps

- **router bgp 39151**
- **neighbor 87.238.112.1 remote-as 65020**

```
neighbor 87.238.112.1 route-map PREFER-THIS-PEER in

Route-map PREFER-THIS-PEER permit 10
  set local-preference 150
```

**The default local preference is 100. Everything higher than this will be preferred.**

**This route map will make the local router prefer routing updates coming from AS 65020 over all other incoming updates.**

# Manipulating path calculations with route maps

- **router bgp 39151**
- **neighbor 87.238.112.1 remote-as 65020**

```
neighbor 87.238.112.1 route-map MAKE-ME-COST-A-LOT out

Route-map MAKE-ME-COST-A-LOT permit 10
  set as-path prepend 39151 39151 39151 39151
```

**Prepending the own AS number to outgoing updates – updates that our router sends to his peers – increases the cost for our peers to route traffic through our own network; in other words, prepends are used when we do NOT want to receive traffic for certain destinations**

# Filtering of BGP routing updates

- **router bgp 39151**
- **neighbor 87.238.112.1 remote-as 65020**

```
neighbor 87.238.112.1 prefix-list ANY-8to24-NET in

ip prefix-list ANY-8to24-NET permit 0.0.0.0/0 ge 8 le 24
```

**In this configuration, only advertised subnets that have at least a /24 size and have a maximum size of /8 are accepted by the local BGP process; other advertised prefixes are filtered out and ignored**

# Dangerous BGP commands

- `clear ip bgp *`

  **This resets all BGP connections with the router and discards the ENTIRE BGP forwarding table. EVERYTHING MUST BE RELEARNED. On a backbone router, this can take hours and while the router is relearning the routes, there is only limited Internet connectivity – if there is any connectivity at all**

```
no router bgp 39151
```

**This command causes the greatest possible damage: It wipes out the entire BGP process AND its configuration**

# Finally: The End