Virtualization and Networking

Why do we want to network computers?

- Initially, to share expensive resources.
    - Not as convenient or as easy as a local asset but the price makes up for it.
- Distributed processing.
- Increased access to information

Network Topologies

- The pattern of connections between the individual machines.
- First broad division is between local area networks (LAN) and wide area networks (WAN).
- WANs
  - Generally speaking are point to point.
  - No addressing needed since there is only one device to read them
  - Since no address no broadcast packet or multicast packet.
  - Frequently WAN links are full duplex so both of the hosts can transmit at the same time.

Network Topologies

- LANs
  - Typically broadcast communications
  - When two hosts are communicating with one another it's across a shared medium.
  - Devices communicating on the LAN have to share the medium.
  - Packets need addresses so broadcast and multicast can be used.

Network Topologies

- Switching blurs the distinction between LANs and WANs.
    - Individual devices are connected directly to ports on the switch.
    - Switch reads the destination from the header and forwards the data to the appropriate port
        - No sharing needed
        - Can run full duplex.

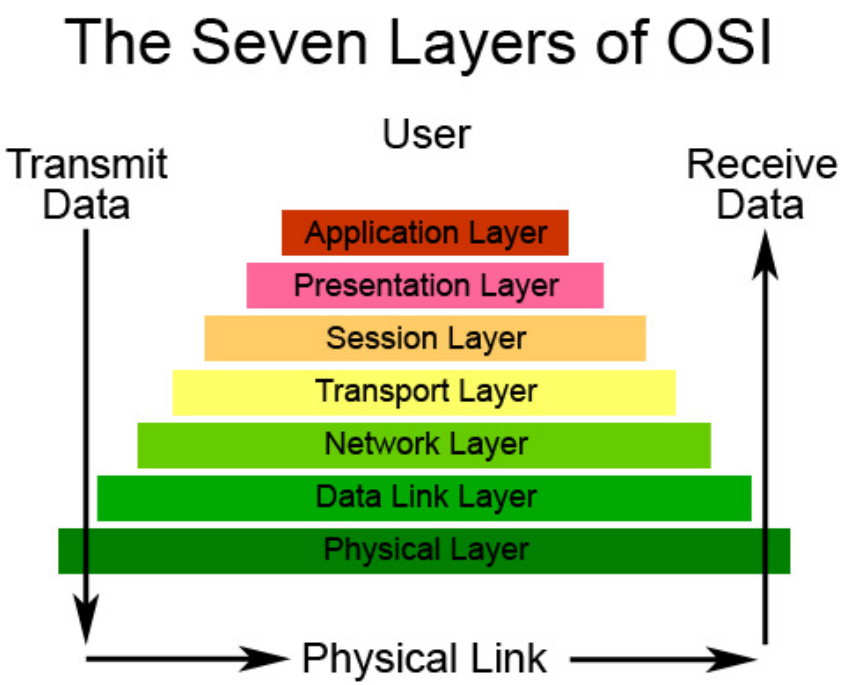
Network Topologies

- Linear
- Hierarchical
- Star
- Ring
- Partly connected mesh
- Fully connected mesh

Models

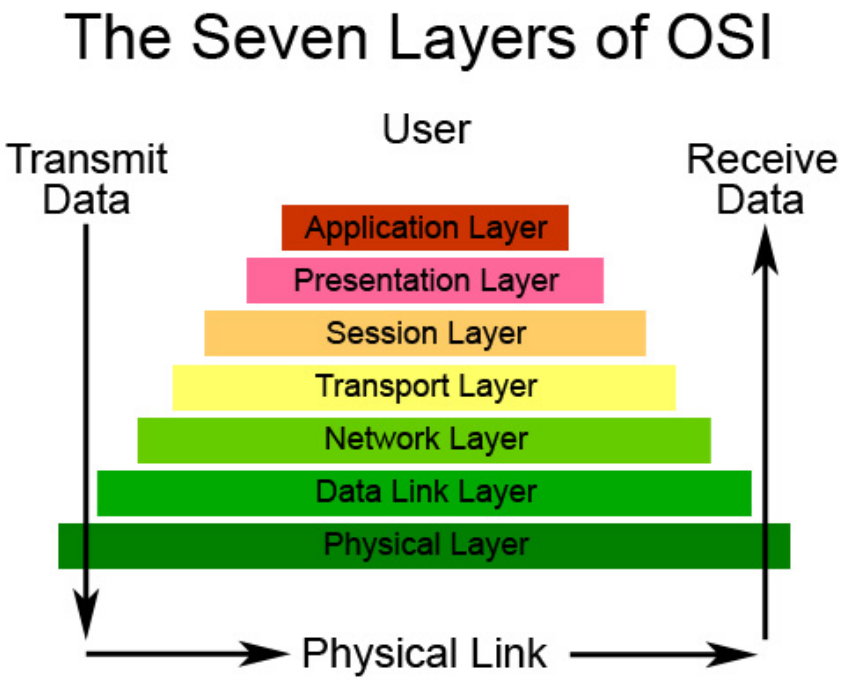
- Models have been created to study and implement networks.
    - Divided into smaller topics which are considered software layers.
    - Each layer provides some service to the next higher layer.
    - Models aren’t perfect
        - Functionality duplicated in layers (security) or crosses boundaries

OSI Model



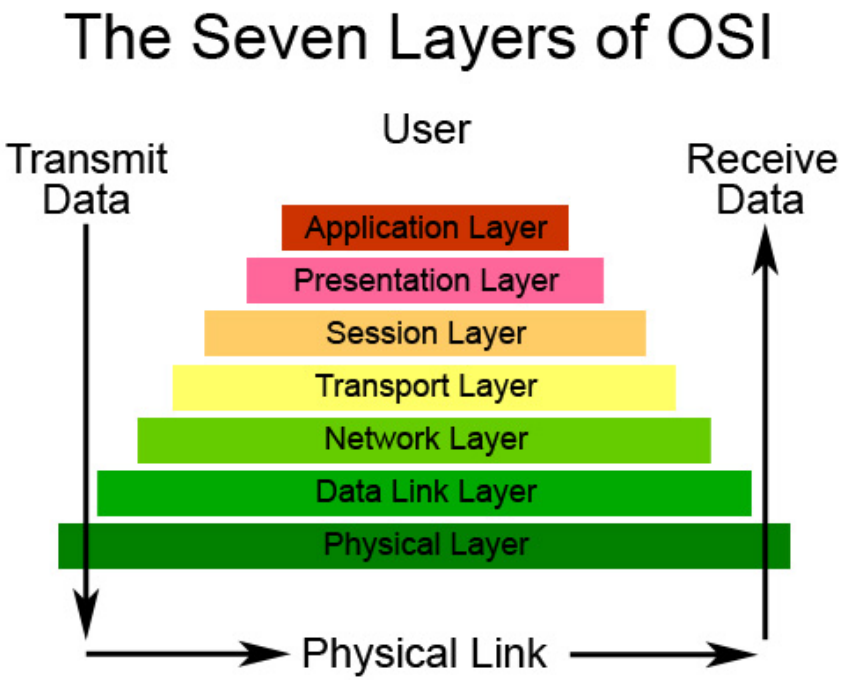
- Most widely known network layer model.
- Developed by the International Standards Organization (ISO).
- Abstract design that described no existing model

OSI Model



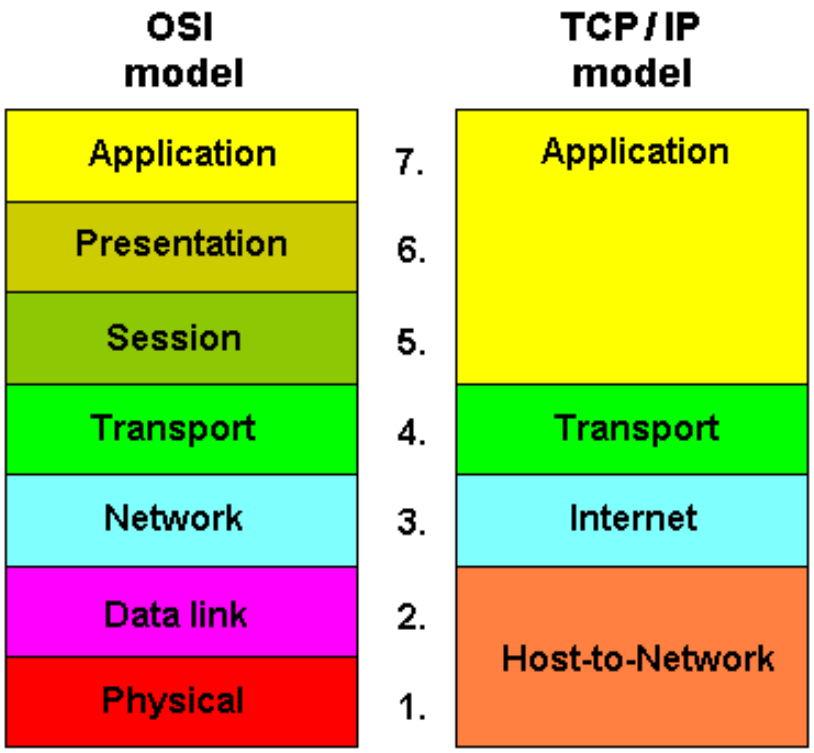
- At one point US government mandated all computer systems purchased by the government implemented this protocol.
  - Never successful and finally abandoned.
- Problems: Presentation and Session layers are never implemented.

OSI Model



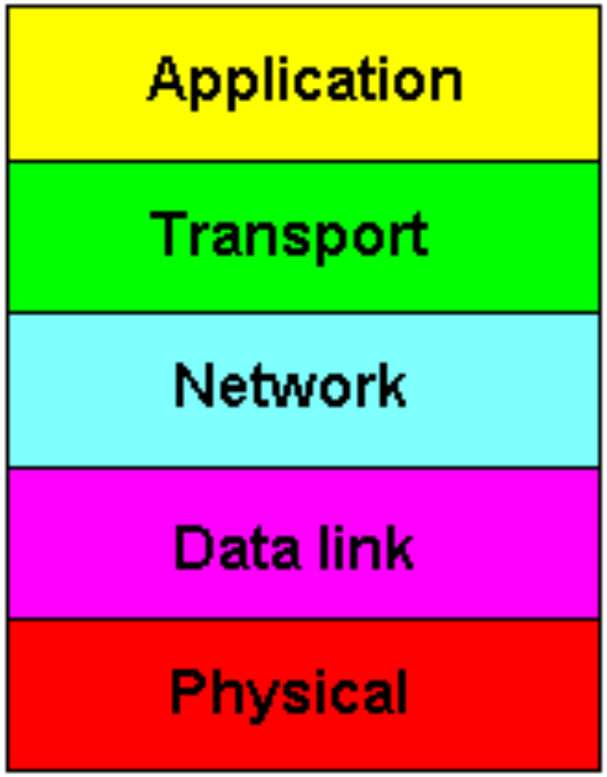
Know this!

TCP/IP Model



- Another model was constructed that described a protocol already in existence, TCP/IP.
- Focused heavily on the upper layers(TCP and IP).
- Ignored the lower layers.
    - Assumes hardware and drivers are commodities to be purchased.

Hybrid Model



- Our book uses a hybrid approach.
- Bottom two layers of the OSI
- Top three of the TCP

Layers: Pros and Cons

- Pro:
  - Small modules are easy to understand, develop, and debug.
  - Can be replaced with newer improved modules.
  - Organizations can specialize in developing different layers.
- Extremely important we have very good definitions of the interfaces, because:
- Con:
  - Many standards

Layers Implementation

- In a networked device there is an entity at each layer that is responsible for the functions of that layer.
- At the physical it's hardware.
- Datalink may be hardware as well.
- Most devices have software for all the other layers.

Physical Layer

- Defines:
  - The actual physical medium used for communicating.
  - The methods and techniques for getting information on and off the medium.
  - Medium may be wire, cable, optical fiber, electromagnetic signal.

Datalink Layer

- Responsible for accessing the shared medium.
- Concerned with packaging information into discrete packets and arbitrating access to the medium.
- Datalink layer devices such as switches or bridges eliminated the need for arbitration in modern networks.

Network layer

- Responsible for routing the information through complex multiple networks with differing physical layer techniques.

Transport Layers

- Creates a reliable connection between two network entities, though not all applications require a connection or a guarantee of reliability.

Transport Layer

- Previous to the Internet there were many different sets of network protocols.
  - Phenomenal success of the Internet changed things.
  - TCP/IP came to dominate the networking landscape.

Transport Layer

- In TCP/IP, IP is the network protocol
- TCP, transmission control protocol is one of the two principle transport layer protocols.
- UDP, user datagram protocol, is the other.
    - Unreliable datagram
- TCP provides “connection-oriented, reliable”, communication.

Application Layer

- A process in one host exchanging information with a process (usually) in another host.

Application Layer

- The entity on one system interacting with the entity on another system.
- Each application will use a specific protocol.
    - Many widely used and have been assigned port numbers.
        - port numbers used by the transport later to determine which application should receive the incoming data.
        - well-known port number - assigned by the IETF to only be used by that application. Valid range 0 - 1023
        - Non well known are 1024-49150
        - 49151 - 65535 - Used for dynamic purposes.

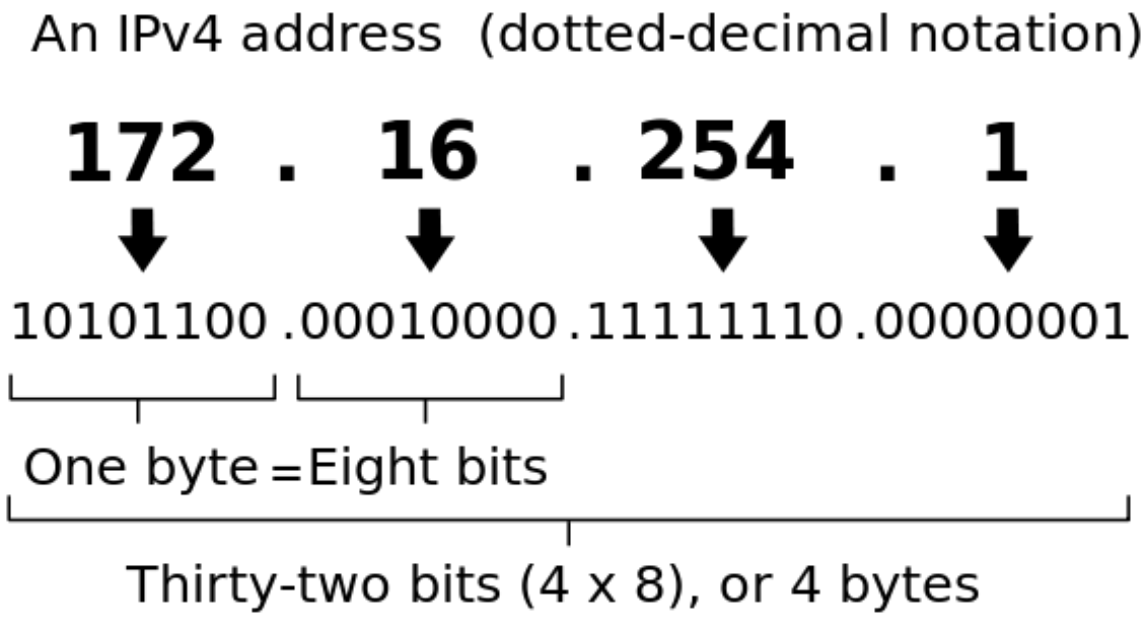
IP Addressing and Routing

- IP Address - a numerical label assigned to each device
- An IP address serves two principal functions: host or network interface identification and location addressing.
- The designers of the Internet Protocol defined an IP address as a 32-bit number
    - Internet Protocol Version 4 (IPv4),
- Due to the enormous growth of the Internet a new version of IP (IPv6), using 128 bits for the address, was developed in 1995.

IP Address

- IPv4 address consists of 32 bits which limits the address space to 4294967296 (2^{32}) possible unique addresses.
    - IPv4 reserves some addresses for special purposes such as private networks (~18 million addresses) or multicast addresses (~270 million addresses).
- IPv4 addresses are represented in dot-decimal notation, which consists of four decimal numbers, each ranging from 0 to 255, separated by dots, e.g., 192.168.20.1. Each part represents a group of 8 bits (octet) of the address.

IP Address



Decomposition of an IPv4 address from dot-decimal notation to its binary value.

IPv4 Addresses

- In the early stages of development of the Internet Protocol, IP addresses were interpreted in two parts: network number portion and host number portion.
    - The highest order octet (most significant eight bits) in an address was designated as the network number and the remaining bits were called the rest field or host identifier and were used for host numbering within a network.
    - Proved inadequate as additional networks developed that were independent of the existing networks already designated by a network number.
- In 1981, the Internet addressing specification was revised with the introduction of classful network architecture.

IPv4 Addresses

- Classful network design allowed for a larger number of individual network assignments and fine-grained subnetwork design. The first three bits of the most significant octet of an IP address were defined as the class of the address.

- Three classes (A, B, and C) were defined for universal unicast addressing. Depending on the class derived, the network identification was based on octet boundary segments of the entire address. Each class used successively additional octets in the network identifier, thus reducing the possible number of hosts in the higher order classes (B and C).

Classful Architecture

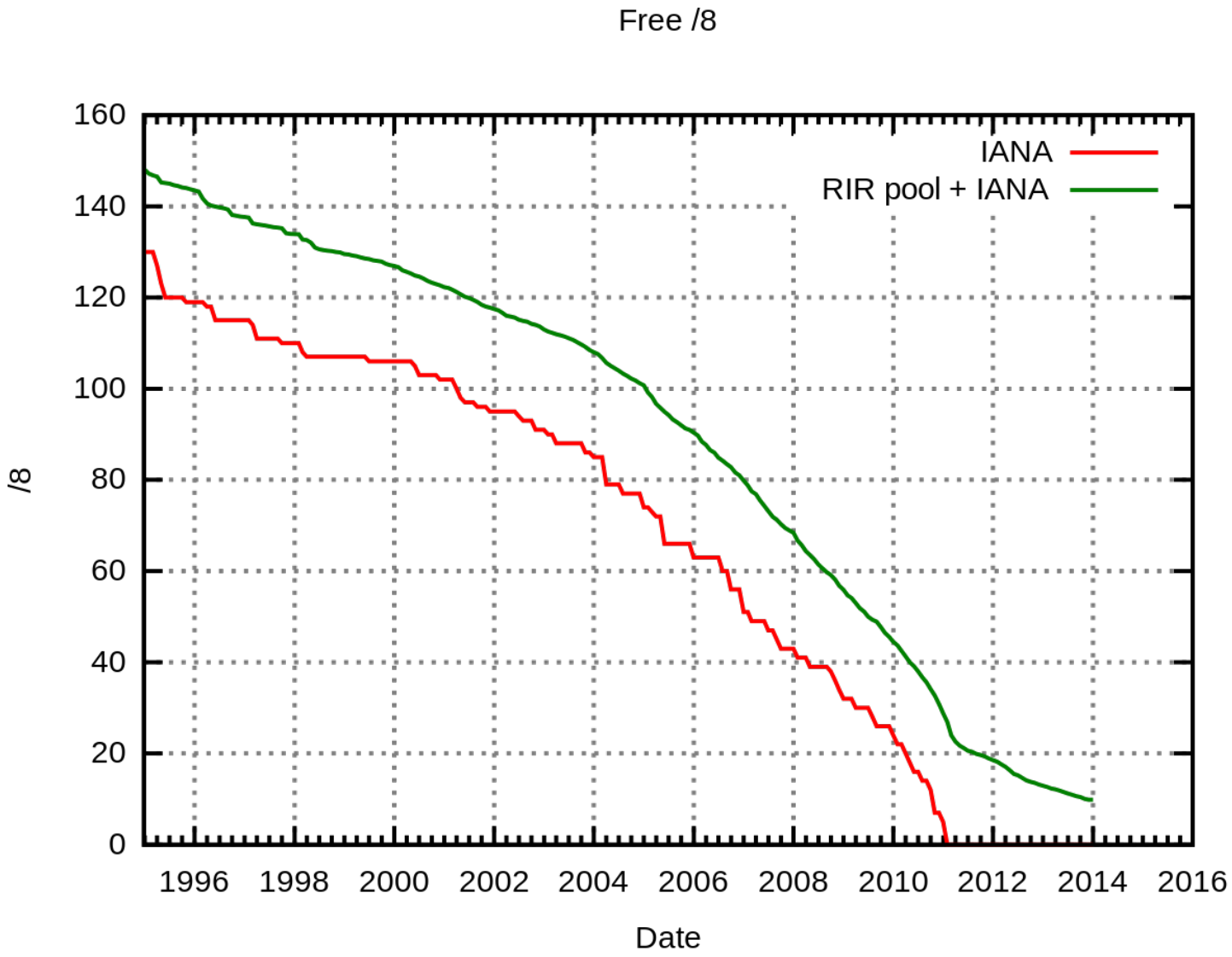
Class	Leading bits	Size of *network number* bit field	Size of *rest* bit field	Number of networks	Addresses per network	Start address	End address
A	0	8	24	128 (2^7)	16,777,216 (2^{24})	`0.0.0.0`	`127.255.255.255`
B	10	16	16	16,384 (2^{14})	65,536 (2^{16})	`128.0.0.0`	`191.255.255.255`
C	110	24	8	2,097,152 (2^{21})	256 (2^8)	`192.0.0.0`	`223.255.255.255`

Historical classful network architecture

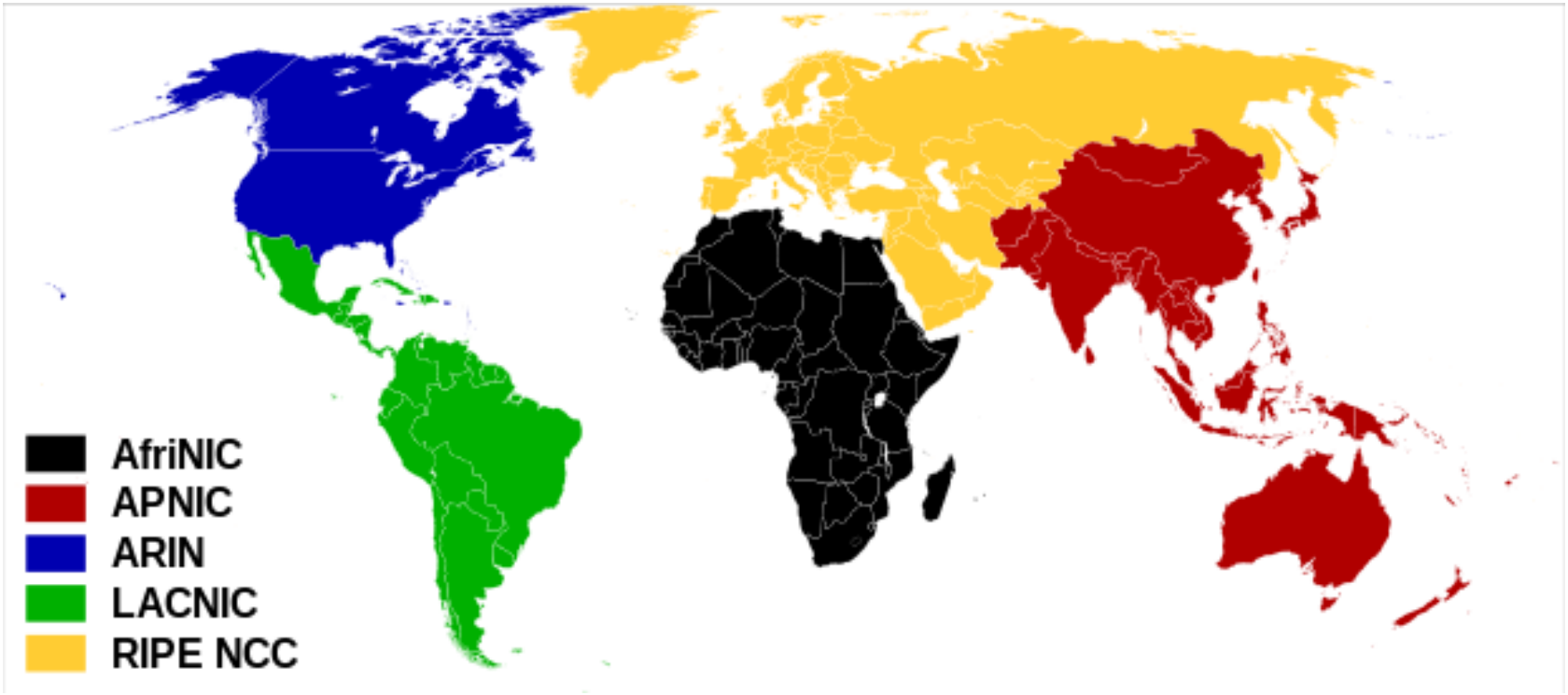
Classless interdomain routing

- In 1993, classful mechanism replaced with CIDR, classless interdomain routing
- Today, technical distinctions of classes of address have gone away.

IPv4 Address Depletion



Regional Internet Registries



Three of the five registries have been depleted:

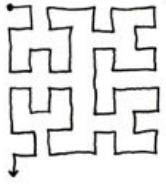
- Asia-Pacific on 15 April 2011
- Europe on 14 September 2012
- Latin America and the Caribbean on 10 June 2014.

MAP OF THE INTERNET

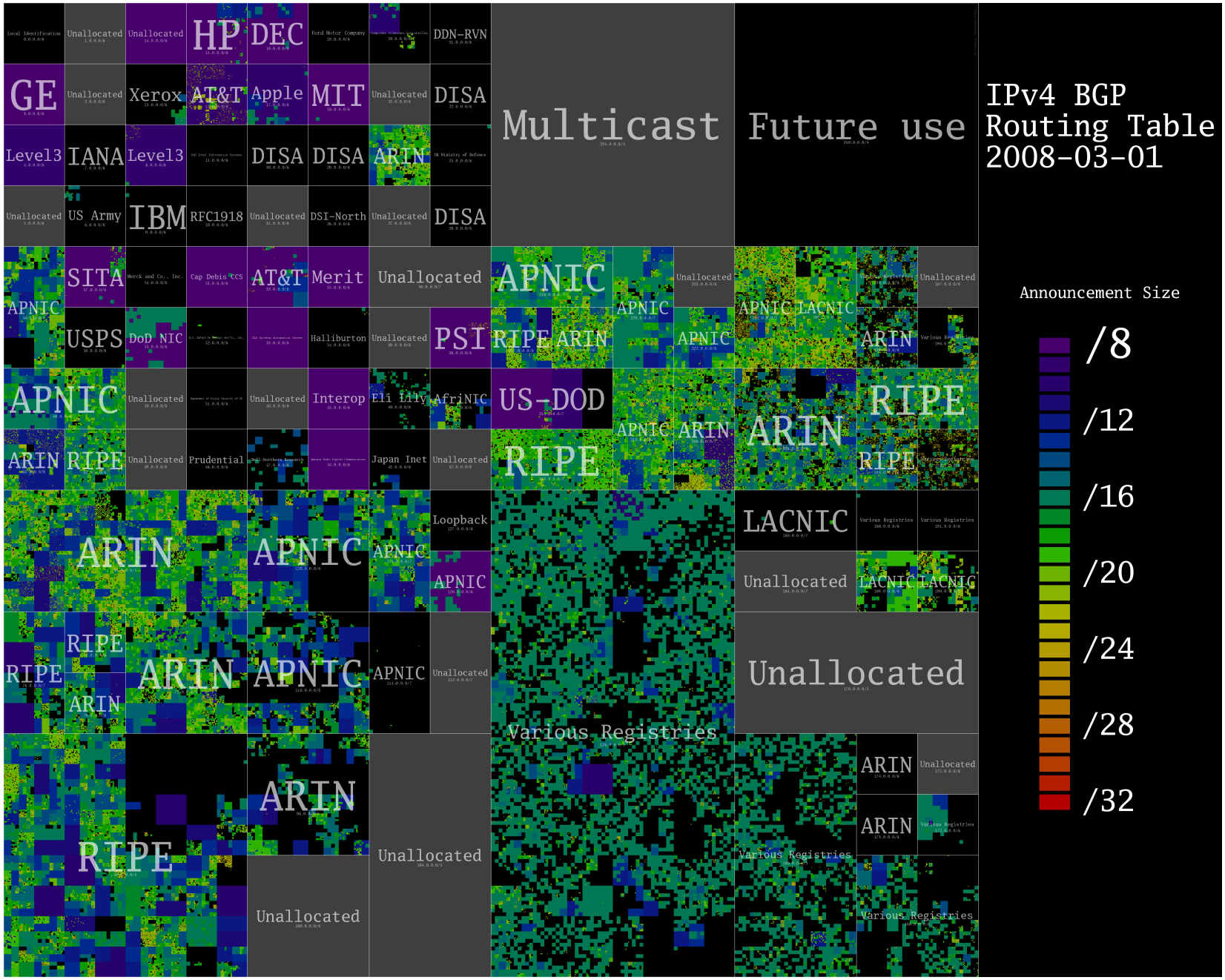
THE IPV4 SPACE, 2006

THIS CHART SHOWS THE IP ADDRESS SPACE ON A PLANE USING A FRACTAL MAPPING WHICH PRESERVES GROUPING -- ANY CONSECUTIVE STRING OF IPs WILL TRANSLATE TO A SINGLE COMPACT, CONTIGUOUS REGION ON THE MAP. EACH OF THE 256 NUMBERED BLOCKS REPRESENTS ONE /8 SUBNET (CONTAINING ALL IPs THAT START WITH THAT NUMBER). THE UPPER LEFT SECTION SHOWS THE BLOCKS SOLD DIRECTLY TO CORPORATIONS AND GOVERNMENTS IN THE 1990's BEFORE THE RIRs TOOK OVER ALLOCATION.

0 1 14 15 16 19 →
3 2 13 12 17 18
4 7 8 11
5 6 9 10



= UNALLOCATED BLOCK

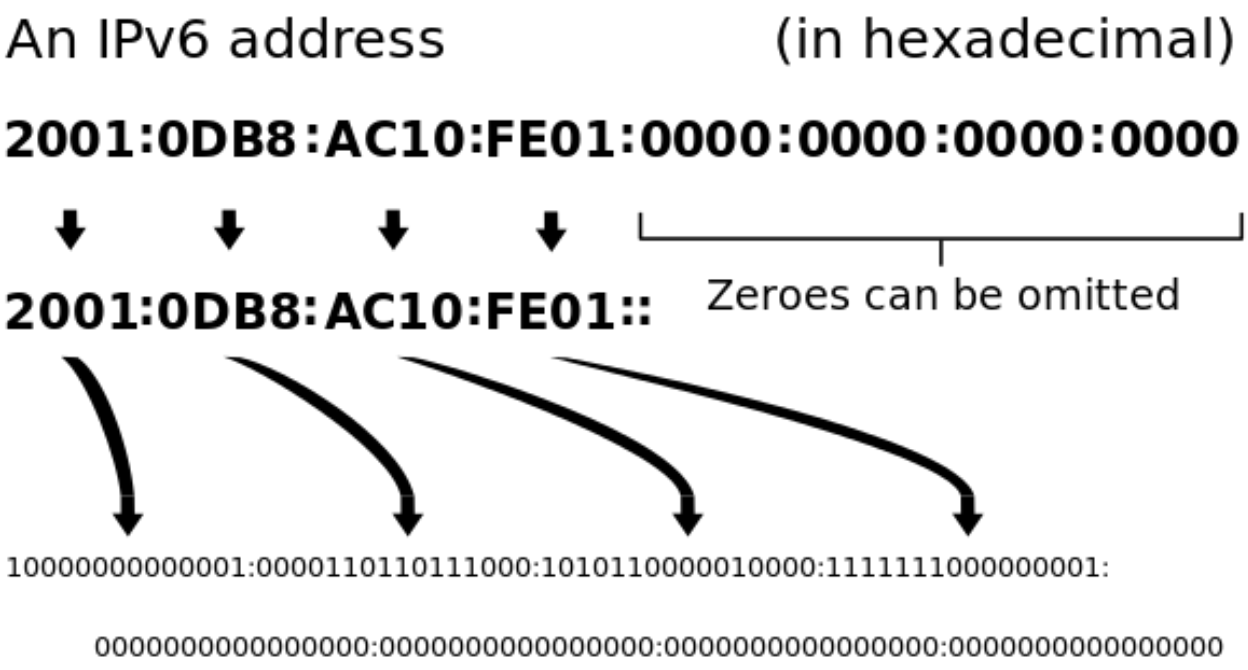


4096 x 4096 image. Each pixel is 256 hosts. Accurate as of 2008

IPv6 Addresses

- The rapid exhaustion of IPv4 address space, prompted the Internet Engineering Task Force (IETF) to expand the addressing capability in the Internet.
  - Address size was increased from 32 to 128 bits or 16 octets.
  - Provides the potential for a maximum of 2^{128}, or about 3.403×10^{38} addresses.

IPv6 Addresses



Routing

- Routers are responsible for delivering IP packets from the source device to the destination device.
  - Routing protocol used between two routers depends on their administrative relationship.
  - Different routing groups: distance vector and link state.

Routing

- Distance Vector
  - RIP (routing information protocol)
  - IGRP (interior gateway routing protocol)
  - RIP2 (routing information protocol version 2)
  - EIGRP (enhanced interior gateway routing protocol)
  - BGP (border gateway protocol)

Routing

- Link State
  - OSPF (open shortest path first)

Routing

- Routers connected in a partial mesh topology.
  - Loss of a link will not partition the network into pieces that can't communicate.
  - Earlier days of the internet the term gateway referred to the class fo device we now call a router.
    - default gateway = default router

DHCP

- Devices can be configured with a predetermined IP address, static
- DHCP (dynamic host configuration protocol)
  - DHCP servers configured with a range of IP addresses.
  - Host that is turned on will broadcast a message looking for the DHCP server.
  - DHCP server will tell the host which IP address to use.
  - Address is leased for some period of time after which it must be renewed.

Name Resolution

- Remembering IP addresses is hard.
- DNS (domain name service) is a protocol that translates from a user-friendly name to IP address
- Fully qualified name - www.uta.edu
    - Parts between the periods known as domains.

Domains

- Domains are organized into a tree structure.
- Top Level Domain (TLD) such as .com, delegated to specific organizations by the Internet Corporation for Assigned Names and Numbers (ICANN), which operates the Internet Assigned Numbers Authority (IANA),

Domains

- Originally, the top-level domain space was organized into three main groups: Countries, Categories, and Multiorganizations.
- IANA today distinguishes the following groups of top-level domains:
- Country-code top-level domains: Two letter domains established for countries or territories
- Internationalized country code top-level domains: ccTLDs in non-Latin character sets (e.g., Arabic or Chinese).
- Generic top-level domains (gTLD): Top-level domains with three or more characters

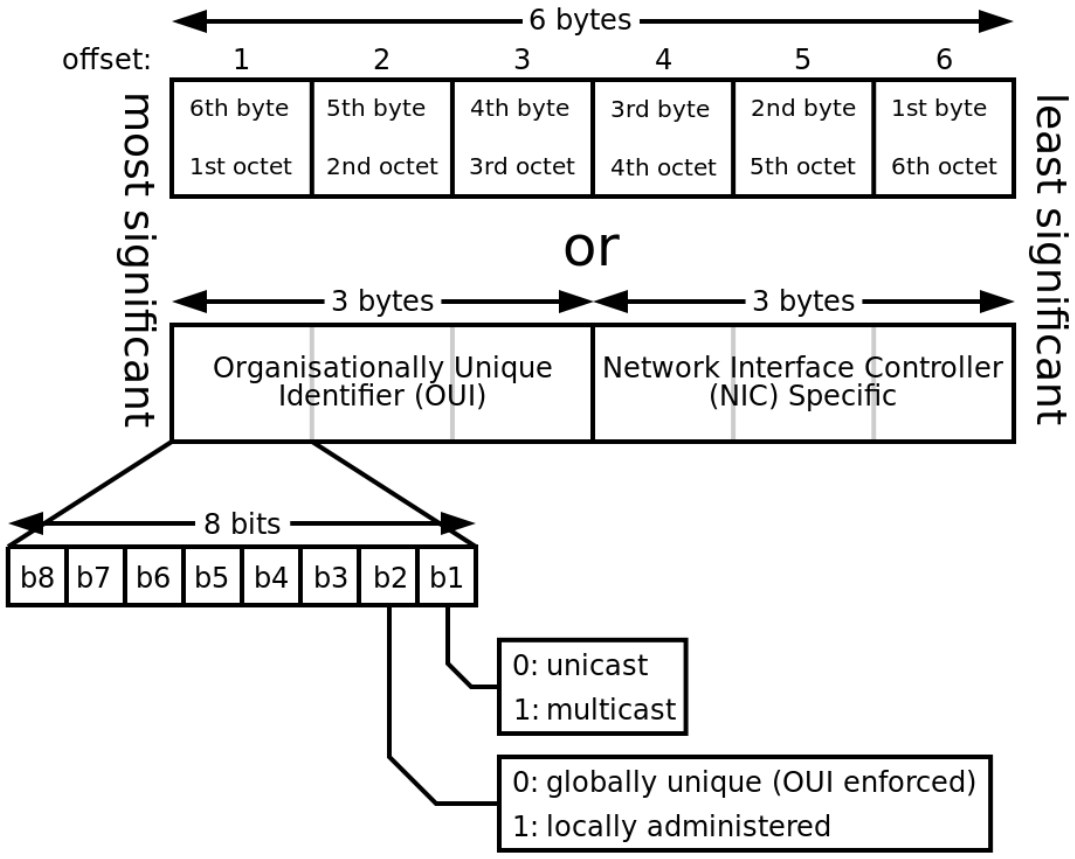
Datalink Layer

- LANs originally had a unique characteristic.
    - Data transmitted in such a way that all the hosts connected to the same link will see every transmission.
        - “multiaccess network”
- Need mechanism to allow hosts to share access. Only read what they should.
    - Media Access Control (MAC)

MAC address

- The original IEEE 802 MAC address comes from the original Xerox Ethernet addressing scheme. This 48-bit address space contains potentially 2^{48} or 281,474,976,710,656 possible MAC addresses.
- Every host is connected to the LAN via network interface card (NIC). Each NIC has a 6 byte identifier
  - Upper three bytes identify the manufacturer
  - Lower three bytes identify the card, uniquely

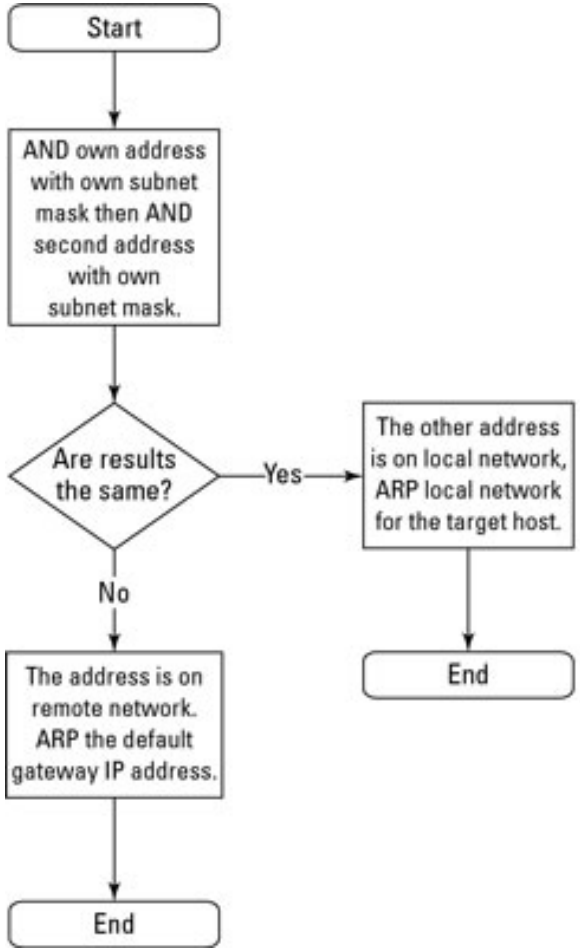
MAC address


6 bytes
offset:
1
2
3
4
5
6
6th byte
1st octet
5th byte
2nd octet
4th byte
3rd octet
3rd byte
4th octet
2nd byte
5th octet
1st byte
6th octet
most significant
least significant
or
3 bytes
3 bytes
Organisationally Unique Identifier (OUI)
Network Interface Controller (NIC) Specific
8 bits
b8
b7
b6
b5
b4
b3
b2
b1
0: unicast
1: multicast
0: globally unique (OUI enforced)
1: locally administered

ARP

- Address resolution protocol (ARP) used to map from IP address to MAC.
  - Host looking for a server will make a broadcast request.
  - All hosts will receive and look into ARP table

ARP

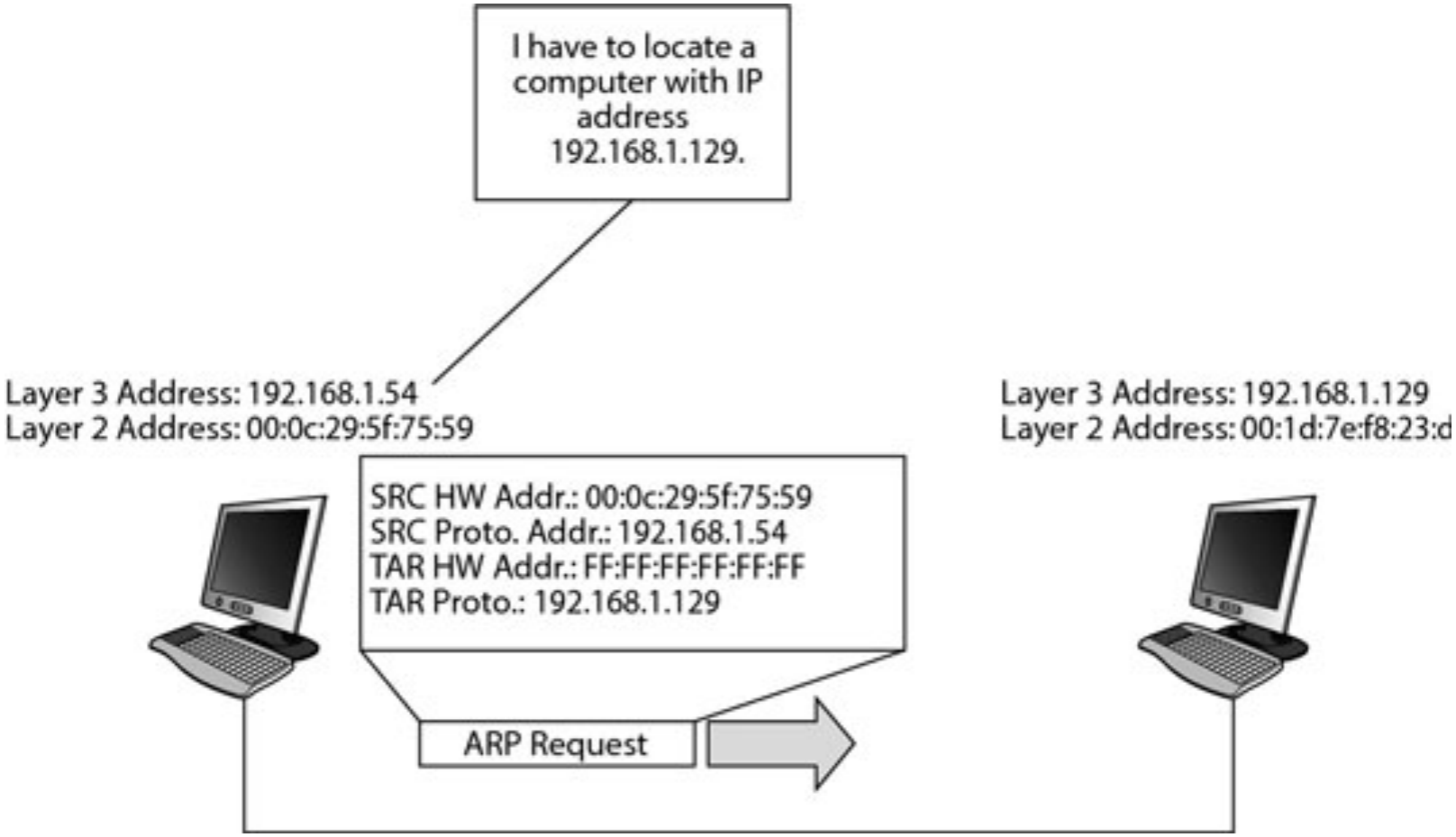

Start
AND own address with own subnet mask then AND second address with own subnet mask.
Are results the same?
Yes
The other address is on local network, ARP local network for the target host.
End
No
The address is on remote network. ARP the default gateway IP address.
End

ARP

1 If the IP address you are trying to communicate with is not in the ARP cache, the address needs to be resolved

2 The data request is placed on hold until the address is resolved and an ARP request is generated and sent onto the network.
All ARP requests have the same basic format: two hardware (or MAC) addresses and two protocol (or IP) addresses (source and target).
The data request includes the sending host's MAC and IP information as well as the IP address of the targeted host. The opcode for this type of packet is 0x0001, denoting that this is a request.

ARP

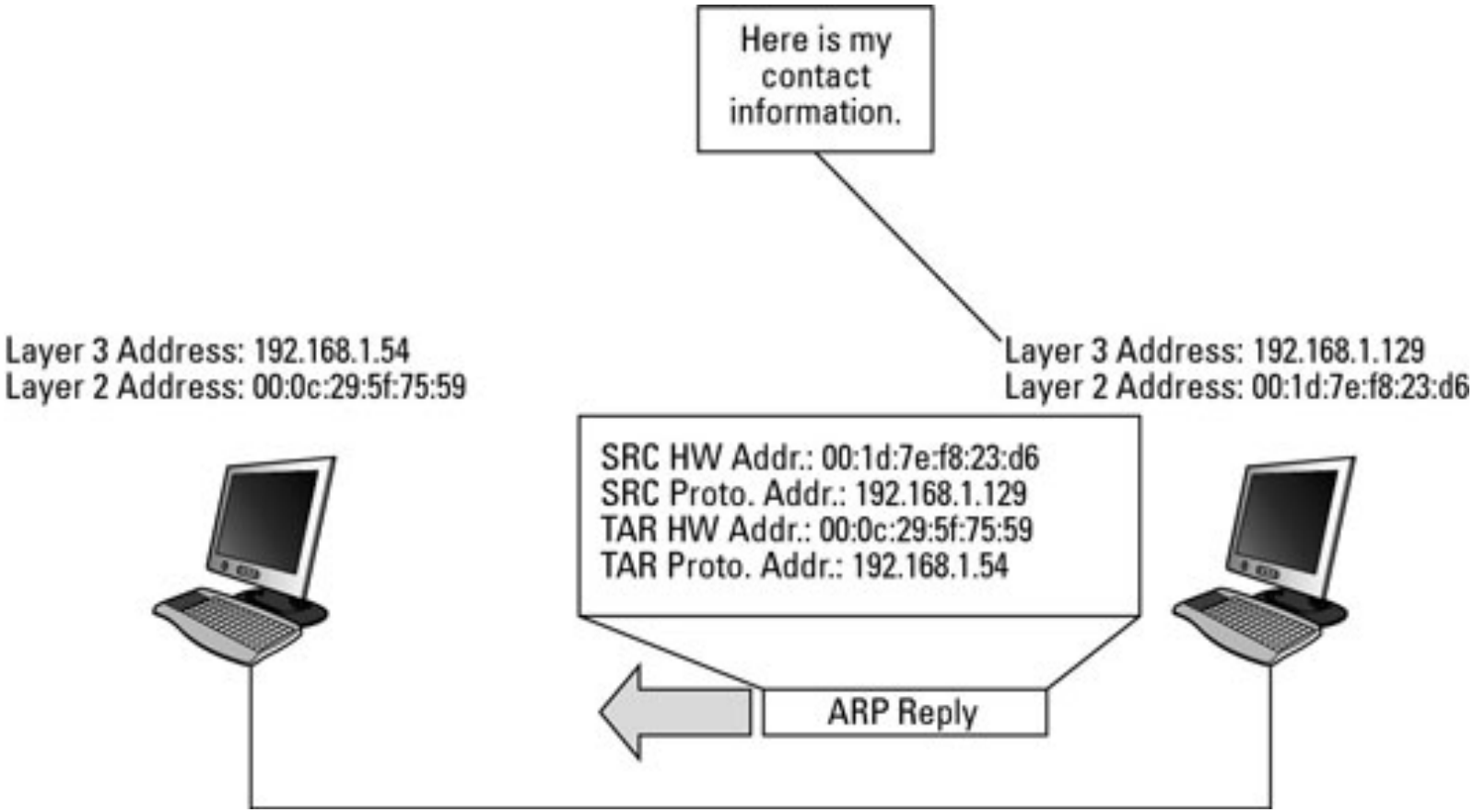

I have to locate a computer with IP address 192.168.1.129.
Layer 3 Address: 192.168.1.54
Layer 2 Address: 00:0c:29:5f:75:59
Layer 3 Address: 192.168.1.129
Layer 2 Address: 00:1d:7e:f8:23:d
SRC HW Addr.: 00:0c:29:5f:75:59
SRC Proto. Addr.: 192.168.1.54
TAR HW Addr.: FF:FF:FF:FF:FF:FF
TAR Proto.: 192.168.1.129
ARP Request

ARP

3 The packet is sent to the local hardware broadcast address, so every computer on the local network segment sees that frame and processes it.
Upon processing the frame and reading the packet information, most computers discard the data because their IP address does not match the one being searched.

4 If by chance, a host does have that address, it records the source MAC and IP address in its own ARP cache, knowing that if someone wants to talk to it, it will likely need to send data shortly, so it then builds its own ARP packet in response.

ARP

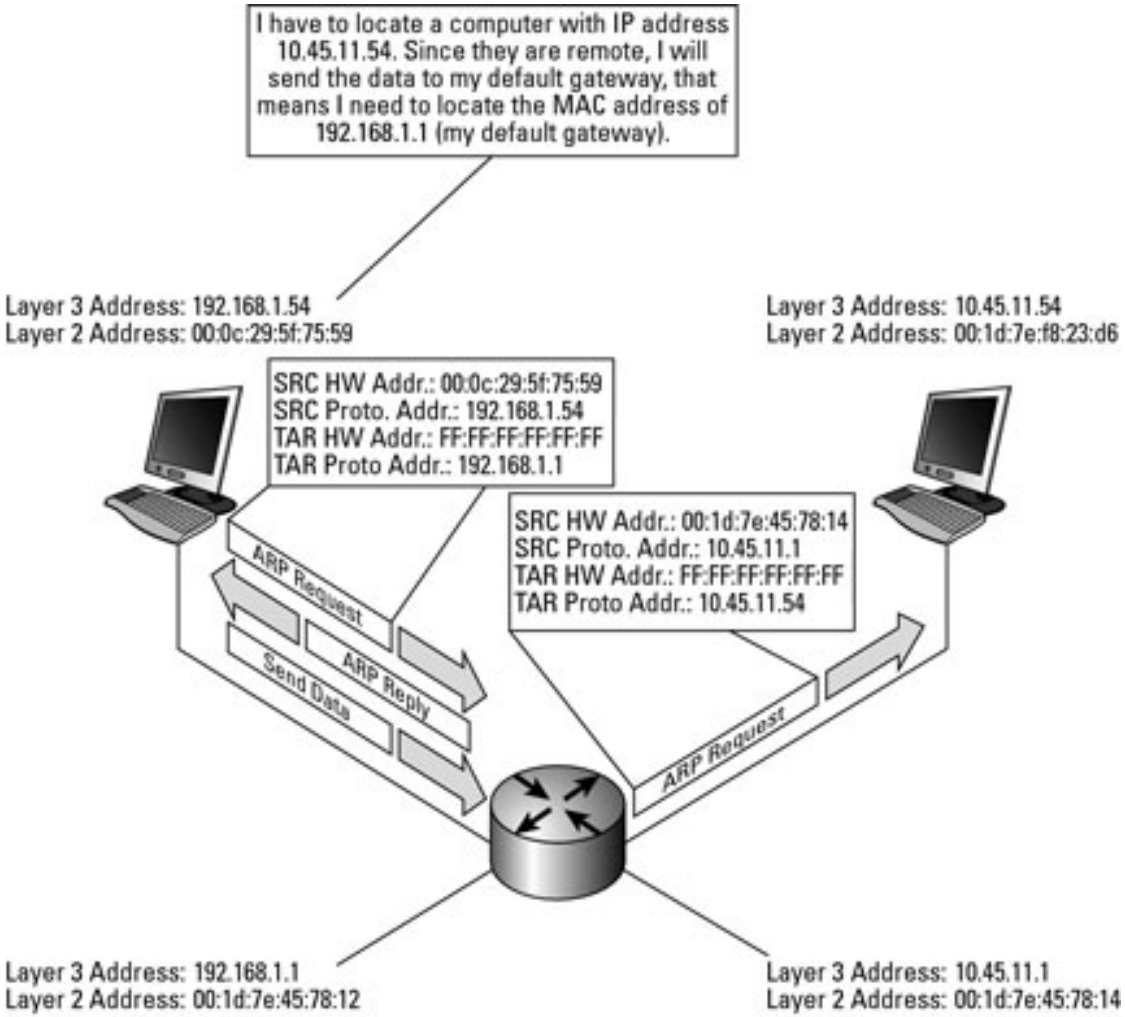

Here is my contact information.
Layer 3 Address: 192.168.1.54
Layer 2 Address: 00:0c:29:5f:75:59
Layer 3 Address: 192.168.1.129
Layer 2 Address: 00:1d:7e:f8:23:d6
SRC HW Addr.: 00:1d:7e:f8:23:d6
SRC Proto. Addr.: 192.168.1.129
TAR HW Addr.: 00:0c:29:5f:75:59
TAR Proto. Addr.: 192.168.1.54
ARP Reply

ARP

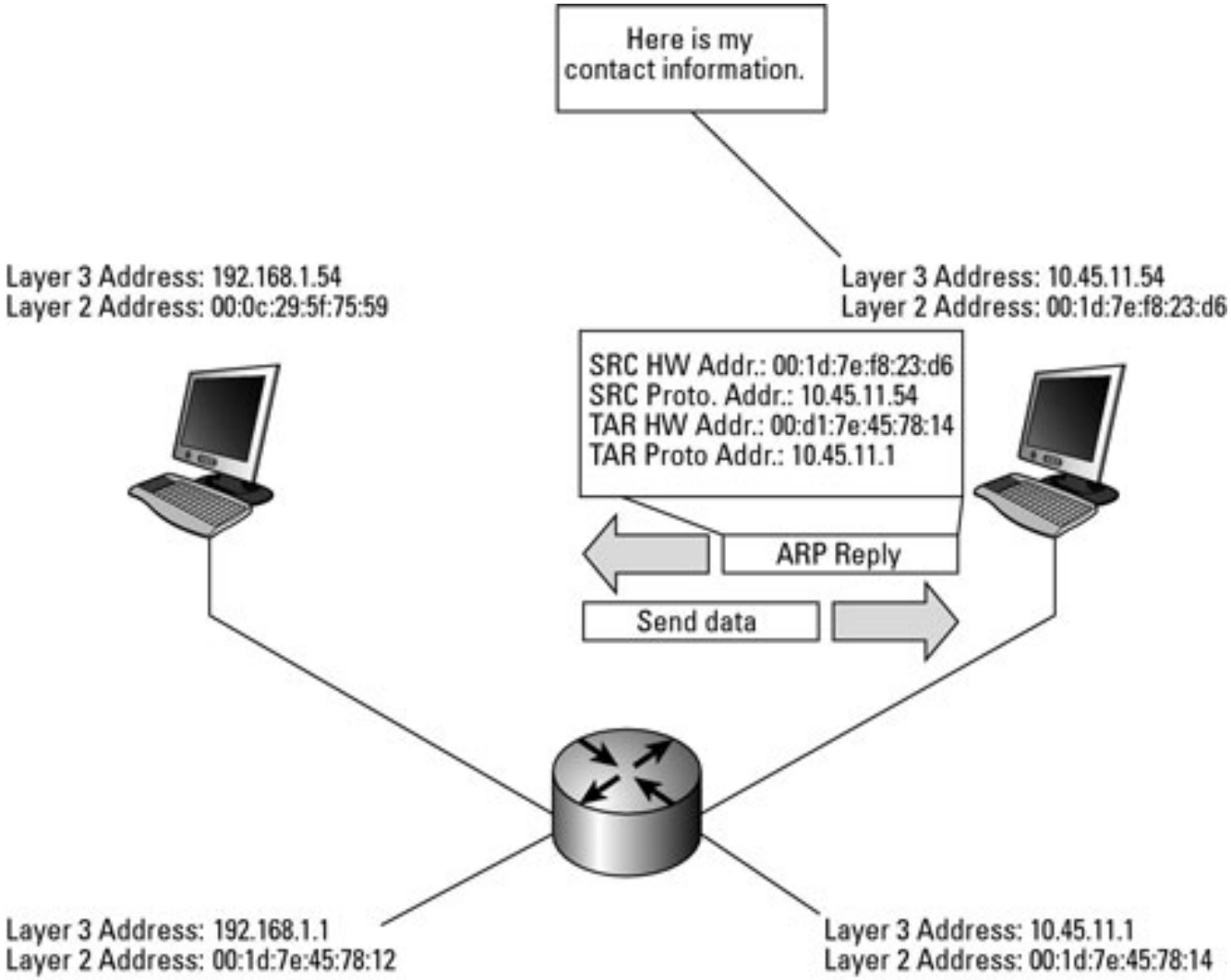
5 With the response sent, the original host sees a frame on the local network segment that is addressed directly to its MAC address; it opens that frame and processes the ARP packet.
The original host then knows the target MAC it needs to send its data to.

6 The original host adds the ARP information to its ARP cache and then releases the data it had placed on hold, sending it to the target MAC address over the local network segment.

ARP


I have to locate a computer with IP address 10.45.11.54. Since they are remote, I will send the data to my default gateway, that means I need to locate the MAC address of 192.168.1.1 (my default gateway).
Layer 3 Address: 192.168.1.54
Layer 2 Address: 00:0c:29:5f:75:59
Layer 3 Address: 10.45.11.54
Layer 2 Address: 00:1d:7e:f8:23:d6
SRC HW Addr.: 00:0c:29:5f:75:59
SRC Proto. Addr.: 192.168.1.54
TAR HW Addr.: FF:FF:FF:FF:FF:FF
TAR Proto Addr.: 192.168.1.1
SRC HW Addr.: 00:1d:7e:45:78:14
SRC Proto. Addr.: 10.45.11.1
TAR HW Addr.: FF:FF:FF:FF:FF:FF
TAR Proto Addr.: 10.45.11.54
ARP Request
ARP Reply
Send Data
ARP Request
Layer 3 Address: 192.168.1.1
Layer 2 Address: 00:1d:7e:45:78:12
Layer 3 Address: 10.45.11.1
Layer 2 Address: 00:1d:7e:45:78:14

ARP


Here is my
contact information.
Layer 3 Address: 192.168.1.54
Layer 2 Address: 00:0c:29:5f:75:59
Layer 3 Address: 10.45.11.54
Layer 2 Address: 00:1d:7e:f8:23:d6
SRC HW Addr.: 00:1d:7e:f8:23:d6
SRC Proto. Addr.: 10.45.11.54
TAR HW Addr.: 00:d1:7e:45:78:14
TAR Proto Addr.: 10.45.11.1
ARP Reply
Send data
Layer 3 Address: 192.168.1.1
Layer 2 Address: 00:1d:7e:45:78:12
Layer 3 Address: 10.45.11.1
Layer 2 Address: 00:1d:7e:45:78:14

Ethernet

- Ethernet relies on the probability that most of the time the network is not busy.
    - If busy, then the sender would wait until it was free and then transmit.
    - If two transmitted at the same time a collision would occur.

Virtualization

» Dates to the 1960's

» VMM (Virtual Machine Monitor) creates the illusion of multiple (virtual) machines on the same physical hardware.

  – Also known as a hypervisor

    • Type 1 hypervisors run on the bare metal

    • Type 2 hypervisors that may make use of an underlying operating system.

– virtualization allows a single computer to host multiple virtual machines

Virtualization

- Advantage of this approach is that a failure in one virtual machine does not bring down any others.
  - Strong Isolation
- BUT, If the server running all the virtual machines fails, the result is even more catastrophic than the crashing of a single dedicated server.
- Only software running in the highest privilege mode is the hypervisor
  - A couple orders of magnitude less lines of code than an OS

Virtualization Advantages

- Checkpointing
- Migrating virtual machines (e.g., for load balancing across multiple servers) is much easier than migrating processes running on a normal operating system.
  - Just move memory and disk images
- Cloud

History

» 1960s IBM experimented with two independently developed hypervisors:
  - SIMMON
  - CP-40
    - Reimplemented as CP-67 to form the control program of CP/CMS, a virtual machine operating system for the IBM System/360 Model 67
    - Reimplemented again and released as VM/370 for the System/370 series in 1972

History

» 1974 two computer scientists at UCLA, Gerald Popek and Robert Goldberg, published“Formal Requirements for Virtualizable Third Generation Architectures”

  – listed exactly what conditions a computer architecture should satisfy in order to support virtualization efficiently

» 1990 researchers at Stanford developed Disco hypervisor

  » Left to form VMWare

    » Binary Translation

Binary Translation

- Basic block: a short, straight-line sequence of instructions that ends with a branch.
  - By definition, a basic block contains no jump, call, trap, return, or other instruction that alters the flow of control, except for the very last instruction
- Prior to executing a basic block, the hypervisor first scans it to see if it contains sensitive instructions and replaces them with a call to a hypervisor procedure that handles them
  - Most code blocks don't contain sensitive instructions

Requirements for Virtualization

- Safety: hypervisor should have full control of virtualized resources.
- Fidelity: behavior of a program on a virtual machine should be identical to same program running on bare hardware.
- Efficiency: much of code in virtual machine should run without intervention by hypervisor.

Safety

» Execute each instruction in an interpreter
  – Bochs
  – Cannot allow the guest operating system to disable interrupts for the entire machine or modify the page-table mappings.
    • Make the OS think it has done that
  – Performance sucks.
    • VMMs try to execute most code natively

Fidelity

- Virtualization has long been a problem on the x86 architecture due to defects in the Intel 386 architecture
  - Carried forward into new CPUs for 20 years in the name of backward compatibility.
    - Sensitive instructions
      - Instructions that do I/O, change the MMU settings
    - Privileged instructions
      - Instructions that cause a trap if executed in user mode

Fidelity

» A machine is virtualizable only if the sensitive instructions are a subset of the privileged instructions

» Some sensitive 386 instructions were ignored if executed in user mode or executed with different behavior.

  – POPF instruction replaces the flags register, which changes the bit that enables/disables interrupts.

    • In user mode, it was ignored.

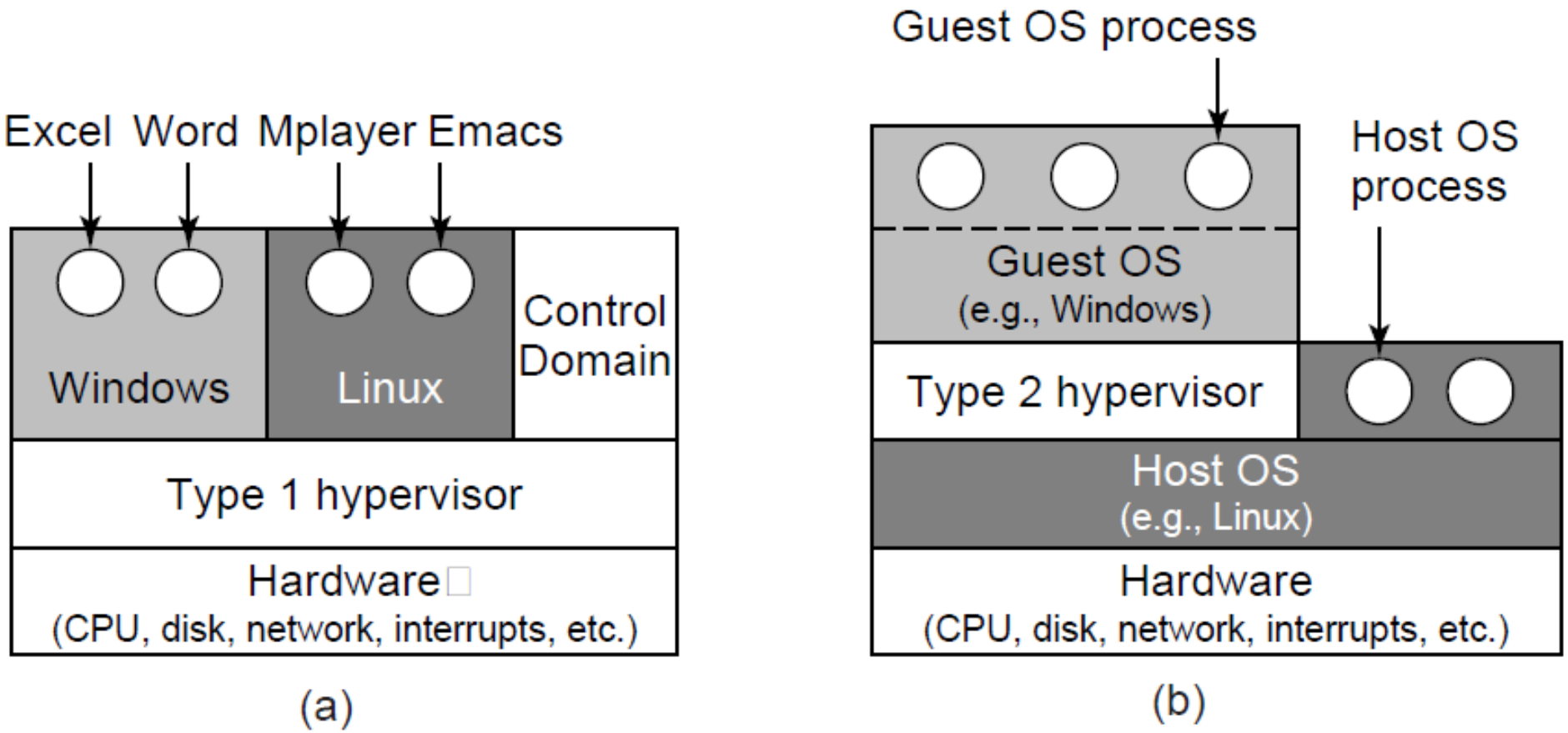
Fidelity

- 2005 Intel released VT (Virtualization Technology); AMD called it SVM (Secure Virtual Machine).
- When a guest operating system is started up in a VT container, it continues to run there until it causes an exception and traps to the hypervisor
- The set of operations that trap is controlled by a hardware bitmap set by the hypervisor.
  - Classical trap-and-emulate becomes possible.

Paravirtualization

- Never even aims to present a virtual machine that looks just like the actual underlying hardware.
- Presents a machine-like software interface that explicitly exposes the fact that it is a virtualized environment.
  - Offers a set of hypercalls, which allow the guest to send explicit requests to the hypervisor
  - Guests use hypercalls for privileged sensitive operations like updating the page tables

Type 1 and Type 2



Location of type 1 and type 2 hypervisors.

Type 1 and Type 2

- Type 1 hypervisor - only program running in ring 0
- Type 2 hypervisor relies on host OS to allocate and schedule resources, very much like a regular process.
  - Also called hosted hypervisors
- Guest Operating System: operating system running on top of the hypervisor
- Host Operating System: operating system running on the hardware

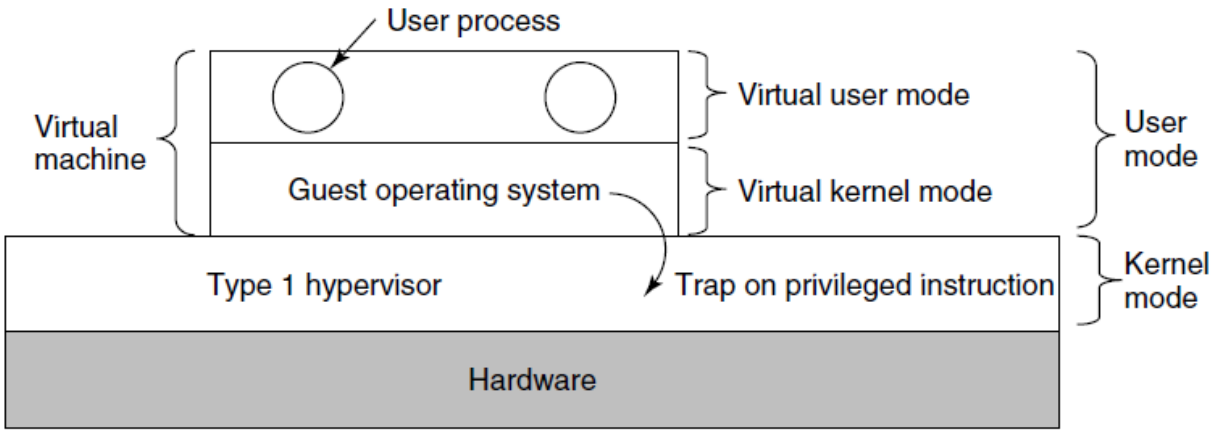
Type 1 and Type 2

Virtualizaton method	Type 1 hypervisor	Type 2 hypervisor
Virtualization without HW support	ESX Server 1.0	VMware Workstation 1
Paravirtualization	Xen 1.0	
Virtualization with HW support	vSphere, Xen, Hyper-V	VMware Fusion, KVM, Parallels
Process virtualization		Wine

Type 1 hypervisors always run on the bare metal

Type 2 hypervisors use the services of an existing host operating system.

Techniques for Efficient Virtualization



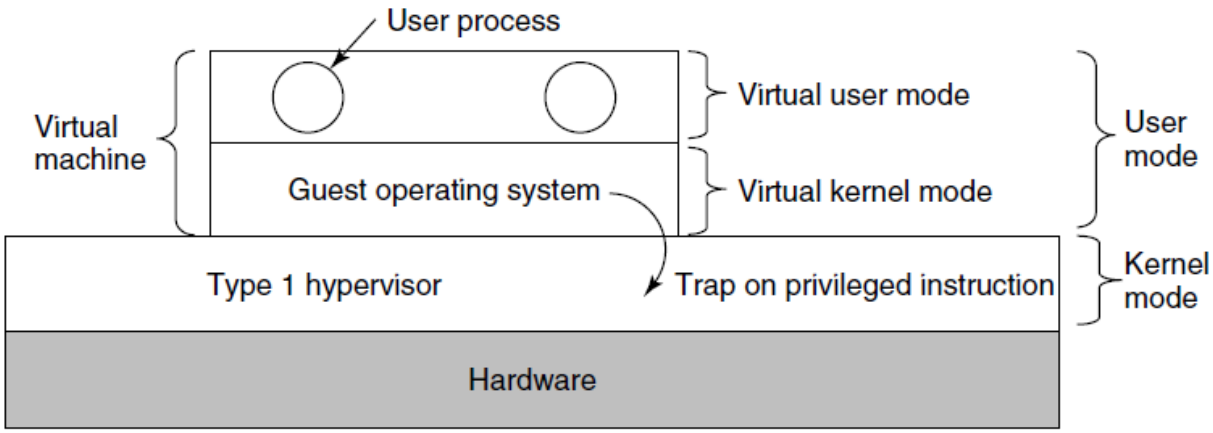
» Virtual machine runs as a user process in user mode, and not allowed to execute sensitive instructions

» Virtual machine runs a guest operating system that thinks it is in kernel mode

– It is not.

– Virtual kernel mode.

Techniques for Efficient Virtualization

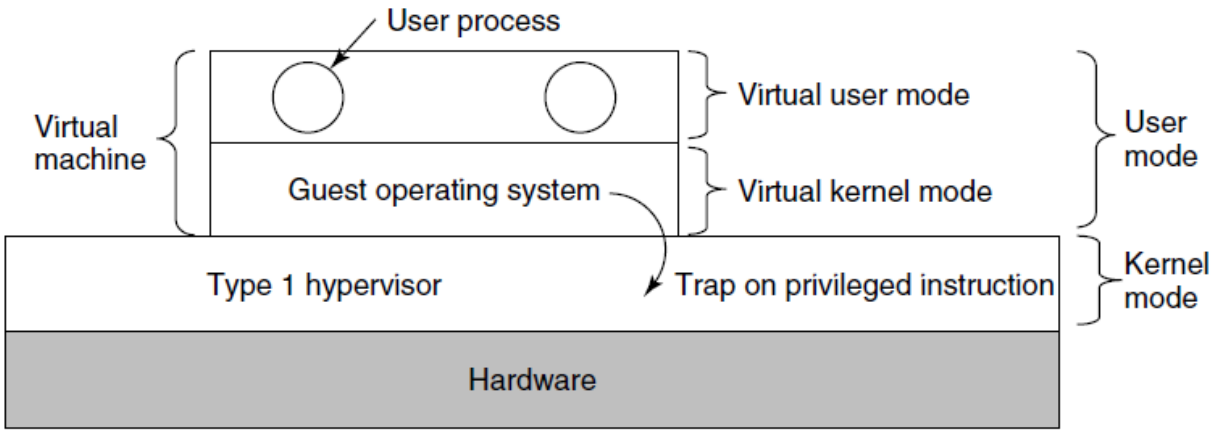


What happens when the guest operating system executes an instruction that is allowed only when the CPU really is in kernel mode?

On CPUs without VT, the instruction fails and the operating system crashes.

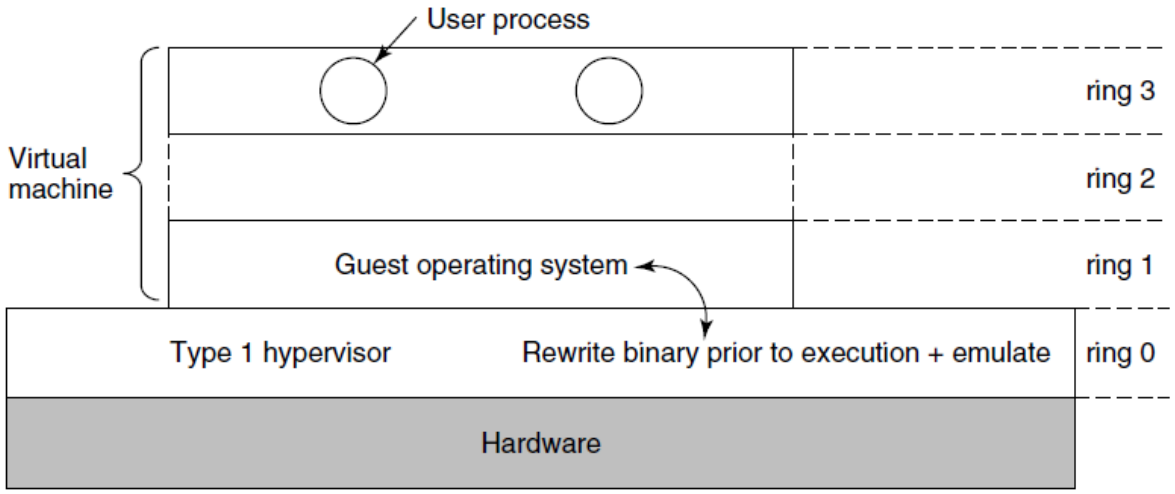
On CPUs with VT, when the guest operating system executes a sensitive instruction, a trap to the hypervisor occurs

Techniques for Efficient Virtualization



- The hypervisor can then inspect the instruction to see if it was issued by the guest operating system in the virtual machine or by a user program in the virtual machine.
  - In the former case, it arranges for the instruction to be carried out;
  - In the latter case, it emulates what the real hardware would do when confronted with a sensitive instruction executed in

Virtualizing the Unvirtualizable



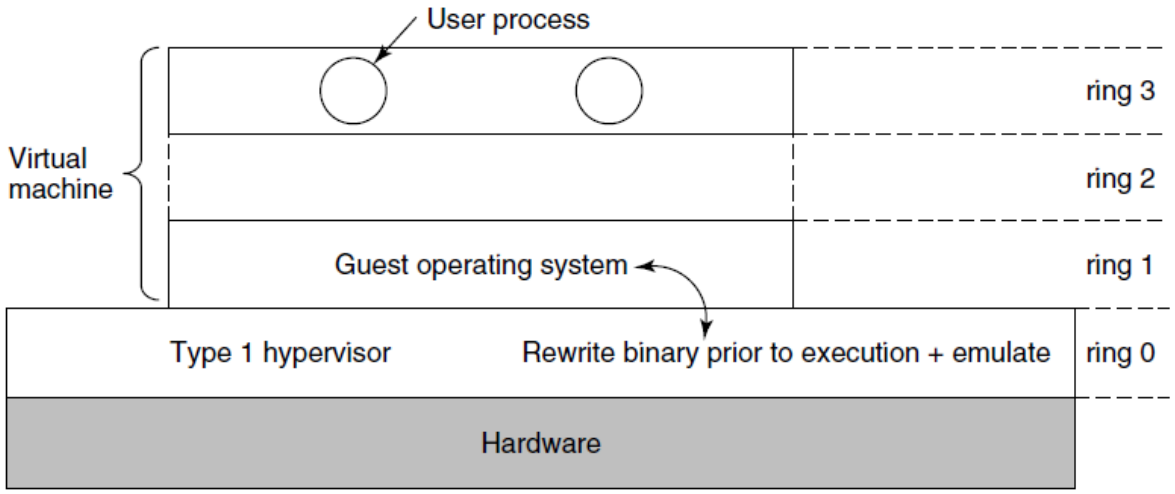
» Virtualizing with VT is straight forward.

» Pre VT:

– Binary translation

– Use rings 1 and 2

Virtualizing the Unvirtualizable



- The kernel is privileged relative to the user processes and any attempt to access kernel memory from a user program leads to an access violation.
- At the same time, the guest operating system's privileged instructions trap to the hypervisor. The hypervisor does some sanity checks and then performs the instructions on the guest's behalf.

World Switch

- Going from a hardware configuration for the host kernel to a configuration for the guest operating system is known as a world switch
  - Interrupts in the guest move the guest kernel mode.
  - Guest kernel expects to be the only kernel in kernel space

Cost of Virtualization

- Trap-and-emulate approach used by VT hardware generates a lot of traps, and traps are very expensive on modern hardware
  - Ruin CPU caches, TLBs, and branch prediction tables internal to the CPU.
  - When sensitive instructions are replaced by calls to hypervisor procedures within the executing process, none of this context-switching overhead is incurred

Cost of Virtualization

» The translated code itself may be either slower or faster than the original code.
  – CLI (Clear Interrupts Instruction)
    • Does not mean the hypervisor should turn them off.
    • Dedicated IF (Interrupt Flag) in the virtual CPU data structure per guest OS.
» Guest operating system modifies its page tables
  – Not cheap.