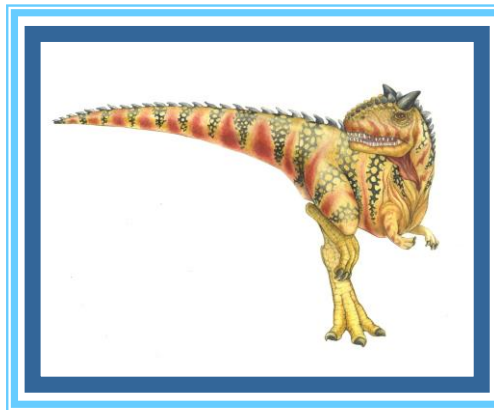# Chapter 1: Introduction to Operating Systems

Objectives

- To define operating system and describe the OS services

- To describe the basic organization of computer systems

- To provide a grand tour of the major components of operating systems

- To give an overview of the many types of computing environments

What is an Operating System?

- A program that acts as an intermediary between a user of a computer and the computer hardware

- Operating system goals:

    - Execute user programs

    - Make the computer system convenient to use

   - Use the computer hardware in an efficient manner

Computer System Structure

- Computer system can be divided into four components:
    - Hardware – provides basic computing resources
        - CPU, memory, I/O devices
    - Operating system
        - Controls and coordinates use of hardware among various applications and users
    - Application programs – define the ways in which the system resources are used to solve the computing problems of the users
        - Word processors, compilers, web browsers, database systems, video games
    - Users
        - People, machines, other computers

Four Components of a Computer System
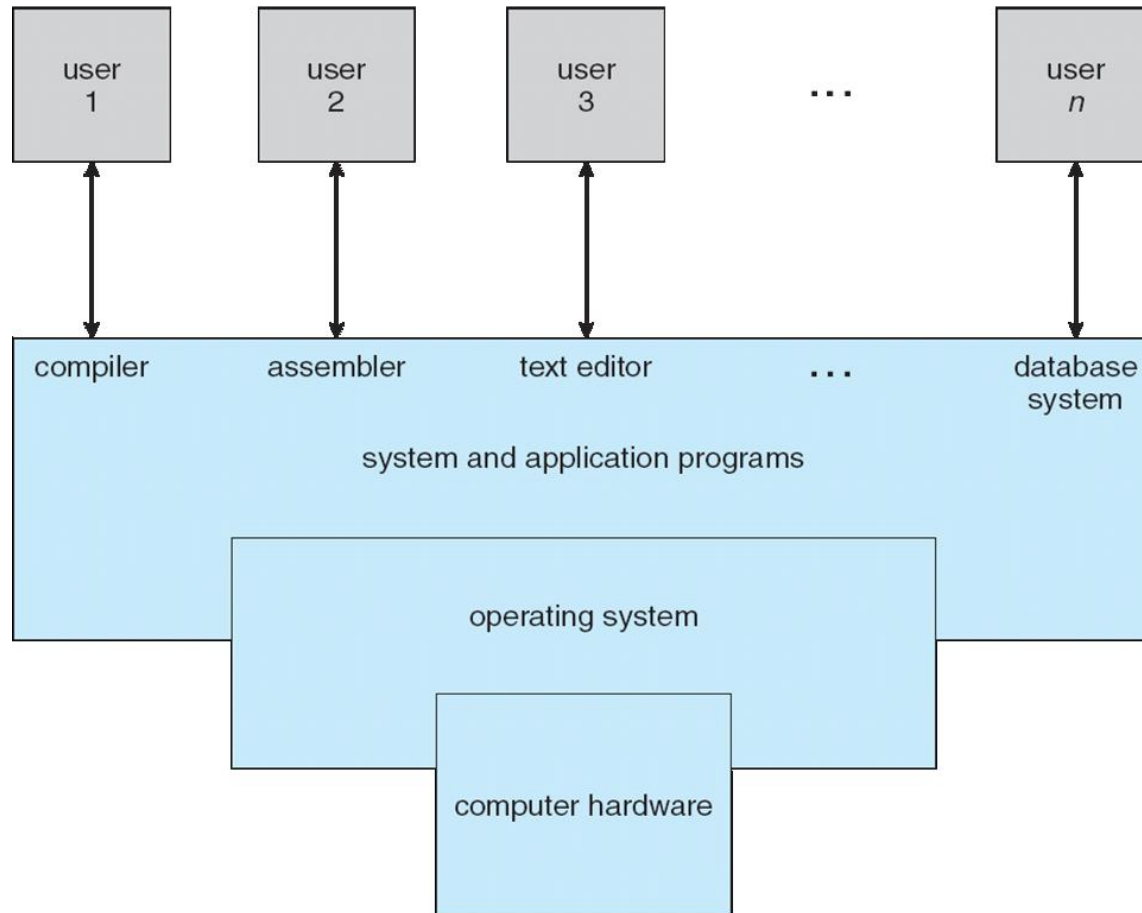
Operating System Definition

- OS is an **interface manager**

  - Make human computer interaction easier

- OS is a **resource allocator**

  - Manages all resources

  - Decides between conflicting requests for efficient and fair resource use

- OS is a **control program**

  - Controls execution of programs to prevent errors and improper use of the computer

  - Multiprogramming system

User Interface

- Operating system provides these facilities for the user:

  - Program creation : editors, debuggers, other development tools.

  - Program execution : load, files, IO operations.

  - Access to IO devices: Read and writes.

  - Controlled access to files: protection mechanisms, abstraction of underlying device.

  - System access: Controls who can access the system.

  - Error detection and response: external, internal, software or hardware error.

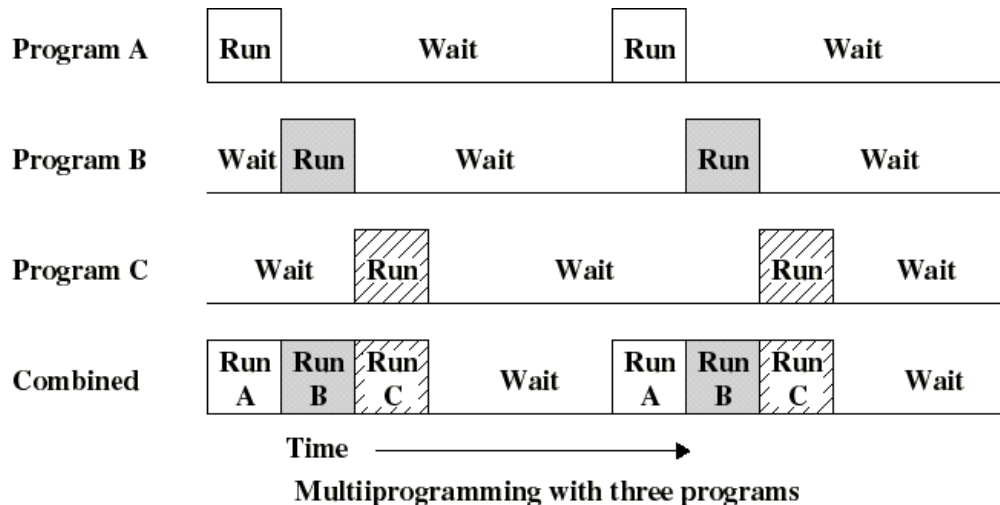 - Accounting: Collect stats., load sharing , for billing purposes

Resource Manager

- Processors : Allocation of processes to processors, preemption, scheduling.

- Memory: Allocation of main memory.

- IO devices : when to access io devices, which ones etc.

- Files: Partitions, space allocation and maintenance.

- Applications, Data, objects.

Multiprogramming

- From uni-programming to multiprogramming systems:

- Multiprogramming systems: batch programs, objective : maximize system (processor) utilization.

- Time sharing systems: Objective is minimize response time. Typical programs are interactive.

- If memory can hold several programs, then CPU can switch to another one whenever a program is awaiting for an I/O to complete

- This is multi-tasking (multi-programming)

Program A	Run	Wait		Run	Wait
Program B	Wait	Run	Wait	Run	Wait
Program C	Wait	Run	Wait	Run	Wait
Combined	Run A	Run B	Run C	Wait	Run A	Run B	Run C	Wait

Time ──────→

Multiiprogramming with three programs

Operating System Services

- Operating systems provide an environment for execution of programs and services to programs and users

- One set of operating-system services provides functions that are helpful to the user:

  - **User interface** - Almost all operating systems have a user interface (**UI**).

    - Varies between **Command-Line (CLI)**, **Graphics User Interface (GUI)**, **Batch**

  - **Program execution** - The system must be able to load a program into memory and to run that program, end execution, either normally or abnormally (indicating error)

  - **I/O operations** - A running program may require I/O, which may involve a file or an I/O device

Operating System Services (Cont.)

- One set of operating-system services provides functions that are helpful to the user (Cont.):

  - **File-system manipulation** - The file system is of particular interest. Programs need to read and write files and directories, create and delete them, search them, list file Information, permission management.

  - **Communications** – Processes may exchange information, on the same computer or between computers over a network
    - Communications may be via shared memory or through message passing (packets moved by the OS)

  - **Error detection** – OS needs to be constantly aware of possible errors
    - May occur in the CPU and memory hardware, in I/O devices, in user program
    - For each type of error, OS should take the appropriate action to ensure correct and consistent computing
    - Debugging facilities can greatly enhance the user's and programmer's abilities to efficiently use the system
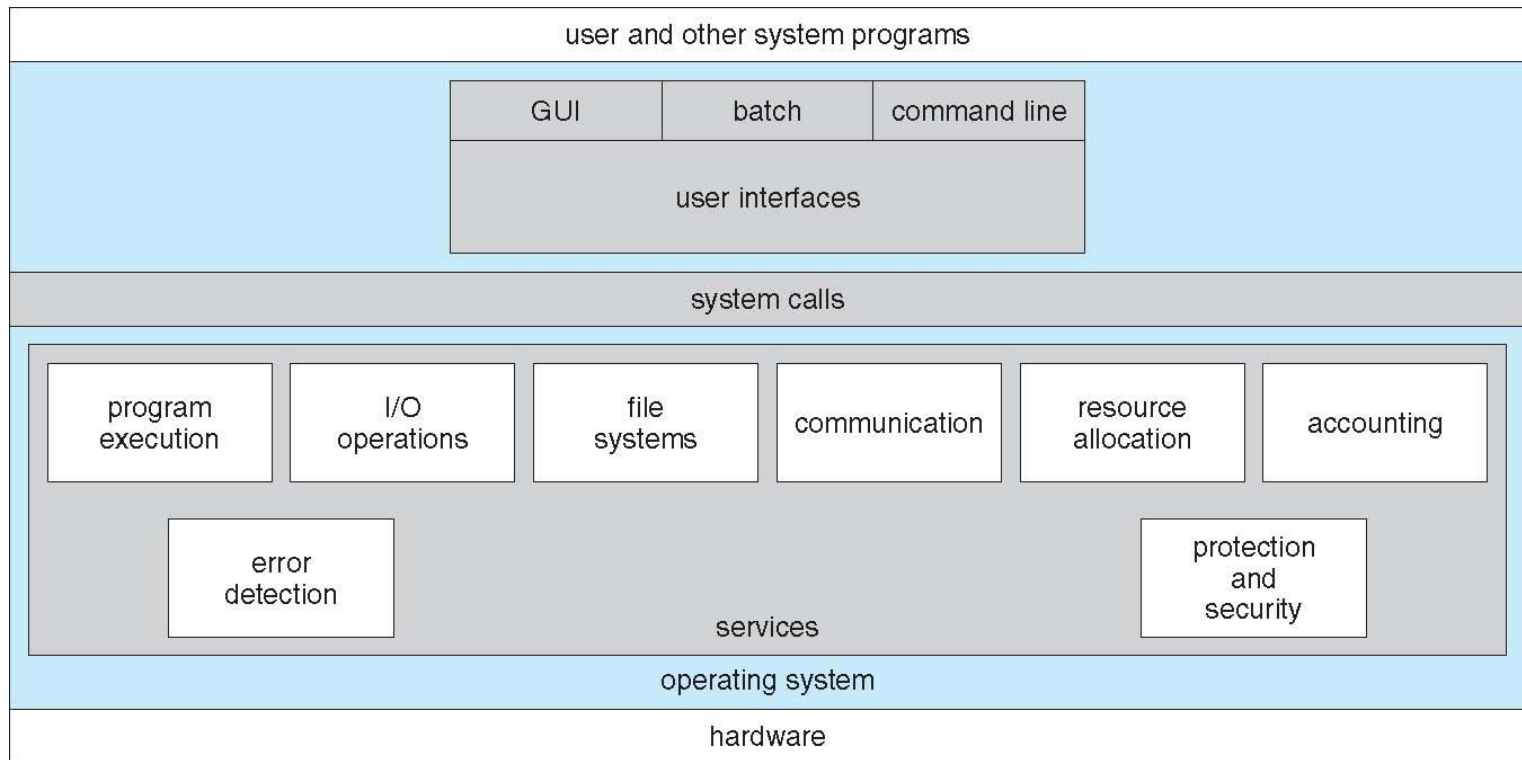
Operating System Services (Cont.)

- Another set of OS functions exists for ensuring the efficient operation of the system itself via resource sharing

  - **Resource allocation -** When multiple users or multiple jobs running concurrently, resources must be allocated to each of them

    - Many types of resources - CPU cycles, main memory, file storage, I/O devices.

  - **Accounting -** To keep track of which users use how much and what kinds of computer resources

  - **Protection and security -** The owners of information stored in a multiuser or networked computer system may want to control use of that information, concurrent processes should not interfere with each other

    - **Protection** involves ensuring that all access to system resources is controlled

    - **Security** of the system from outsiders requires user authentication, extends to defending external I/O devices from invalid access attempts

A View of Operating System Services

user and other system programs

GUI	batch	command line
user interfaces		

system calls

program execution	I/O operations	file systems	communication	resource allocation	accounting

| error detection | | | | protection and security |

services

operating system

hardware

Process Concept

- A process is a program in execution. It is a unit of work within the system. Program is a *passive entity*, process is an *active entity*

- Multiprogramming, time-sharing and real-time transaction systems lead to the refinement of the concept of process.

- A process can be defined by its attributes and behaviors.

- When instances of this process type co-exist we have concurrent processing.

- Issues in concurrent processing : synchronization, mutual exclusion, deadlock, communication.

How a Modern Computer Works
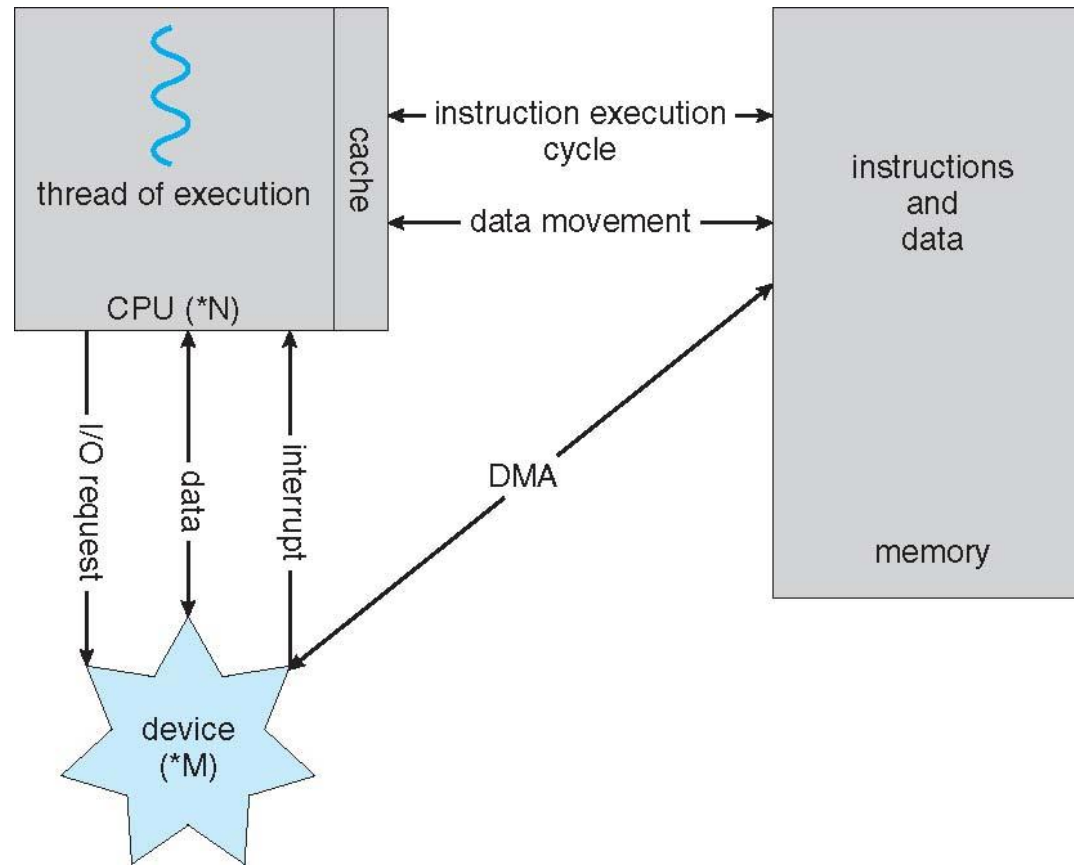
A von Neumann architecture

Process Management

- Process needs resources to accomplish its task
    - CPU, memory, I/O, files
    - Initialization data
- Process termination requires reclaim of any reusable resources
- Single-threaded process has one **program counter** specifying location of next instruction to execute
    - Process executes instructions sequentially, one at a time, until completion
- Multi-threaded process has one program counter per thread
- Typically system has many processes, some user, some operating system running concurrently on one or more CPUs
    - Concurrency by multiplexing the CPUs among the processes / threads

Process Management Activities

The operating system is responsible for the following activities in connection with process management:

- Creating and deleting both user and system processes

- Suspending and resuming processes

- Providing mechanisms for process synchronization
  - priority of the process
  - the event (if any) after which the process is waiting

- Providing mechanisms for process communication

- Providing mechanisms for deadlock handling

Memory Management

- To execute a program all (or part) of the instructions must be in memory

- All (or part) of the data that is needed by the program must be in memory.

- Memory management determines what is in memory and when
  - Optimizing CPU utilization and computer response to users

- Memory management activities
  - Keeping track of which parts of memory are currently being used and by whom
  - Deciding which processes (or parts thereof) and data to move into and out of memory
  - Allocating and deallocating memory space as needed

- Virtual memory and file system facilities together satisfy all these requirements.

- Virtual memory allows programs to address the memory from a logical point of view without regard to the amount of main memory available.

Scheduling and Resource Management

- Scheduling and resource management is an Operations Research (OR) problem.

- Goals : Efficient use of resources, satisfy the service time requested by a process, say, in a real-time system and of course, fairness.

- Short-term and long-term scheduling.

- Queuing is one of the basic operations associated with scheduling. Interrupt is another important concept in the context of scheduling.

- Differential responsiveness
  - differentiate between different classes of jobs

- Fairness
  - give equal and fair access to all processes of the same class

- Efficiency
  - maximize throughput, minimize response time, and accommodate as many users as possible

Storage Management

- OS provides uniform, logical view of information storage
    - Abstracts physical properties to logical storage unit - **file**
    - Each medium is controlled by device (i.e., disk drive, tape drive)
        - ▸ Varying properties include access speed, capacity, data-transfer rate, access method (sequential or random)

- File-System management
    - Files usually organized into directories
    - Access control on most systems to determine who can access what
    - OS activities include
        - ▸ Creating and deleting files and directories
        - ▸ Primitives to manipulate files and directories
        - ▸ Mapping files onto secondary storage
        - ▸ Backup files onto stable (non-volatile) storage media

Performance of Various Levels of Storage

Level	1	2	3	4	5
Name	registers	cache	main memory	solid state disk	magnetic disk
Typical size	< 1 KB	< 16MB	< 64GB	< 1 TB	< 10 TB
Implementation technology	custom memory with multiple ports CMOS	on-chip or off-chip CMOS SRAM	CMOS SRAM	flash memory	magnetic disk
Access time (ns)	0.25 - 0.5	0.5 - 25	80 - 250	25,000 - 50,000	5,000,000
Bandwidth (MB/sec)	20,000 - 100,000	5,000 - 10,000	1,000 - 5,000	500	20 - 150
Managed by	compiler	hardware	operating system	operating system	operating system
Backed by	cache	main memory	disk	disk	disk or tape

Movement between levels of storage hierarchy can be explicit or implicit

I/O Subsystem

- One purpose of OS is to hide peculiarities of hardware devices from the user

- I/O subsystem responsible for

    - Memory management of I/O including buffering (storing data temporarily while it is being transferred), caching (storing parts of data in faster storage for performance)

    - General device-driver interface

    - Drivers for specific hardware devices
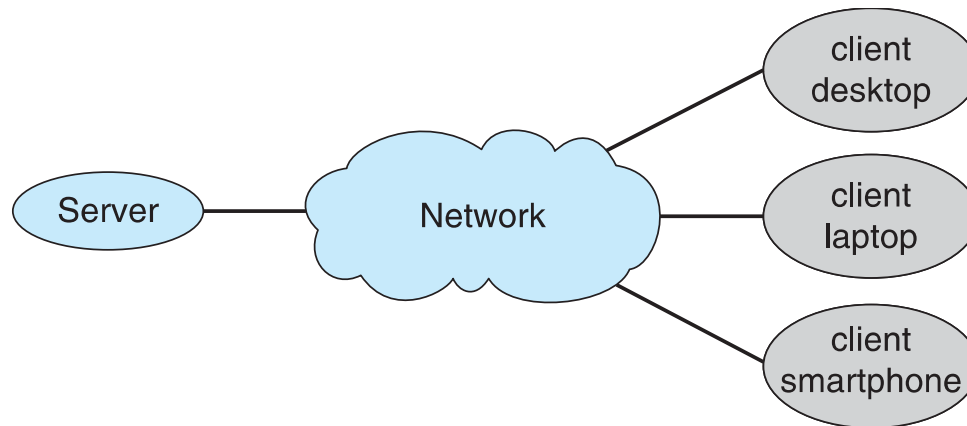
Computing Environments - Traditional

- ☐ Stand-alone general purpose machines

- ☐ But blurred as most systems interconnect with others (i.e., the Internet)

- ☐ **Portals** provide web access to internal systems

- ☐ **Network computers** (**thin clients**) are like Web terminals

- ☐ Mobile computers interconnect via **wireless networks**

- ☐ Networking becoming ubiquitous – even home systems use **firewalls** to protect home computers from Internet attacks

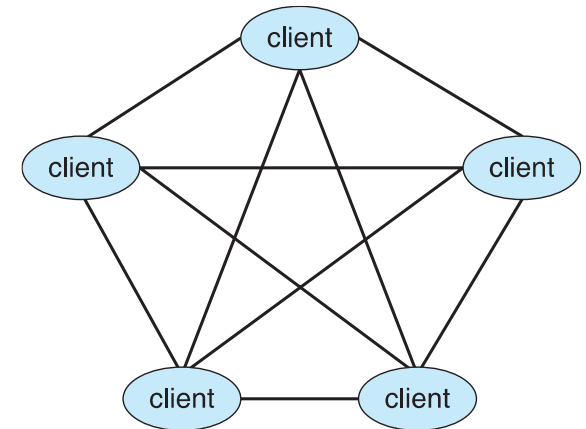# Computing Environments – Client-Server

- Client-Server Computing
  - Dumb terminals supplanted by smart PCs
  - Many systems now **servers**, responding to requests generated by **clients**
    - ▸ **Compute-server system** provides an interface to client to request services (i.e., database)
    - ▸ **File-server system** provides interface for clients to store and retrieve files

Computing Environments - Peer-to-Peer

- Another model of distributed system

- P2P does not distinguish clients and servers

  - Instead all nodes are considered peers

  - May each act as client, server or both

  - Node must join P2P network

    - Registers its service with central lookup service on network, or

    - Broadcast request for service and respond to requests for service via *discovery protocol*

  - Examples include Napster and Gnutella, **Voice over IP** (**VoIP**) such as Skype

Computing Environments – Distributed

- Distributed computiing
  - Collection of separate, possibly heterogeneous, systems networked together
    - **Network** is a communications path, **TCP/IP** most common
      - **Local Area Network** (**LAN**)
      - **Wide Area Network** (**WAN**)
      - **Metropolitan Area Network** (**MAN**)
      - **Personal Area Network** (**PAN**)
  - **Network Operating System** provides features between systems across network
    - Communication scheme allows systems to exchange messages
    - Illusion of a single system

Computing Environments - Virtualization

- Allows operating systems to run applications within other OSes

  - Vast and growing industry

- **Emulation** used when source CPU type different from target type (i.e. PowerPC to Intel x86)

  - Generally slowest method

  - When computer language not compiled to native code – **Interpretation**

- **Virtualization** – OS natively compiled for CPU, running **guest** OSes also natively compiled

  - Consider VMware running WinXP guests, each running applications, all on native WinXP **host** OS

  - **VMM** (virtual machine Manager) provides virtualization services
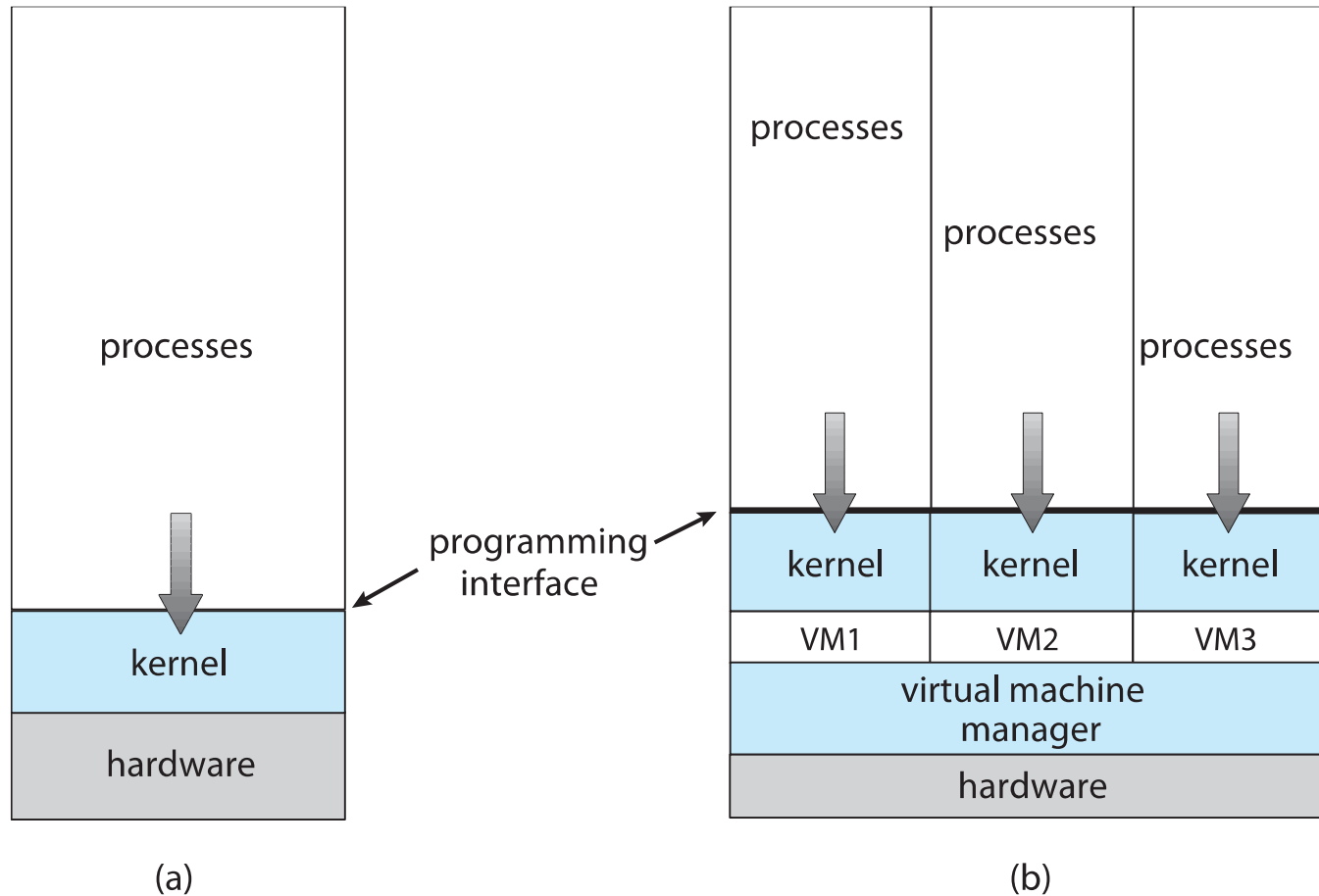
Computing Environments - Virtualization

- Use cases involve laptops and desktops running multiple OSes for exploration or compatibility

  - Apple laptop running Mac OS X host, Windows as a guest

  - Developing apps for multiple OSes without having multiple systems

  - QA testing applications without having multiple systems

  - Executing and managing compute environments within data centers

- VMM can run natively, in which case they are also the host

  - There is no general purpose host then (VMware ESX and Citrix XenServer)

Computing Environments - Virtualization

programming interface

processes

kernel

hardware

(a)

processes   processes   processes

kernel   kernel   kernel

VM1   VM2   VM3

virtual machine manager

hardware

(b)

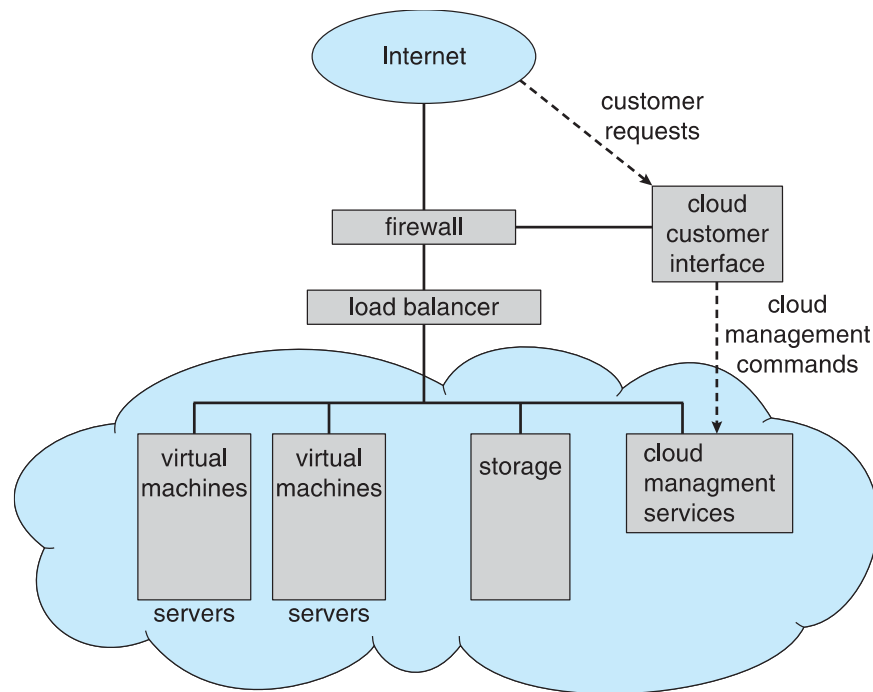# Computing Environments – Cloud Computing

- Delivers computing, storage, even apps as a service across a network
- Logical extension of virtualization because it uses virtualization as the base for it functionality.
  - Amazon **EC2** has thousands of servers, millions of virtual machines, petabytes of storage available across the Internet, pay based on usage
- Many types
  - **Public cloud** – available via Internet to anyone willing to pay
  - **Private cloud** – run by a company for the company's own use
  - **Hybrid cloud** – includes both public and private cloud components
  - Software as a Service (**SaaS**) – one or more applications available via the Internet (i.e., word processor)
  - Platform as a Service (**PaaS**) – software stack ready for application use via the Internet (i.e., a database server)
  - Infrastructure as a Service (**IaaS**) – servers or storage available over Internet (i.e., storage available for backup use)

Computing Environments – Cloud Computing

☐ Cloud computing environments composed of traditional OSes, plus VMMs, plus cloud management tools

  ☐ Internet connectivity requires security like firewalls

  ☐ Load balancers spread traffic across multiple applications

Computing Environments - Mobile

- Handheld smartphones, tablets, etc

- What is the functional difference between them and a "traditional" laptop?

- Extra feature – more OS features (GPS, gyroscope)

- Allows new types of apps like *augmented reality*

- Use IEEE 802.11 wireless, or cellular data networks for connectivity

- Leaders are **Apple iOS** and **Google Android**

- *New features can lead to new applications and services*

  - Discussion, what new features, and what new applications?

  - E.g., accelerometer, finger print, facial recognition, etc…