

# **OPERATING SYSTEMS**



# UNIT -1



# COMPUTER SYSTEM AND OPERATING SYSTEM OVERVIEW

## OVER VIEW OF OPERATING SYSTEM

### 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:

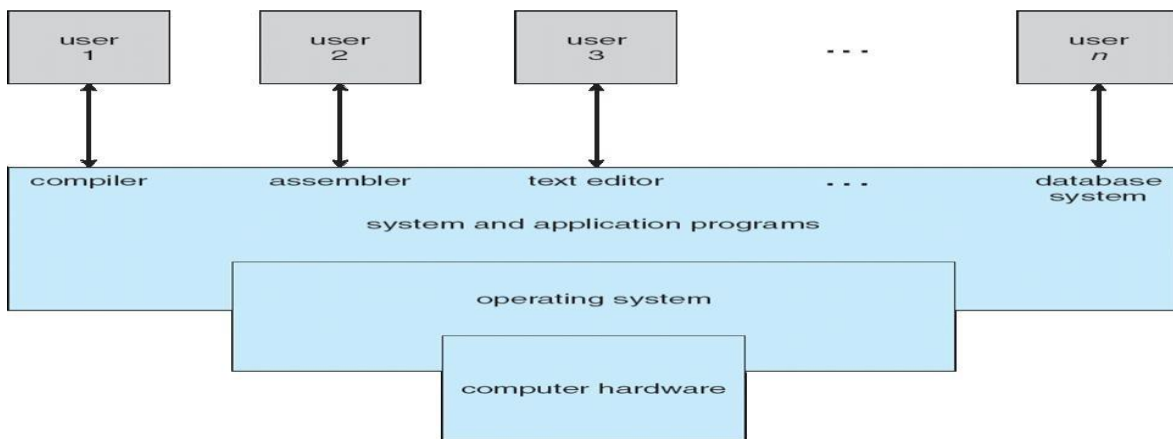
- Execute user programs and make solving user problems easier
- Make the computer system convenient to use
- Use the computer hardware in an efficient manner

### 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 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
- No universally accepted definition
- Everything a vendor ships when you order an operating system” is good approximation
  - But varies wildly

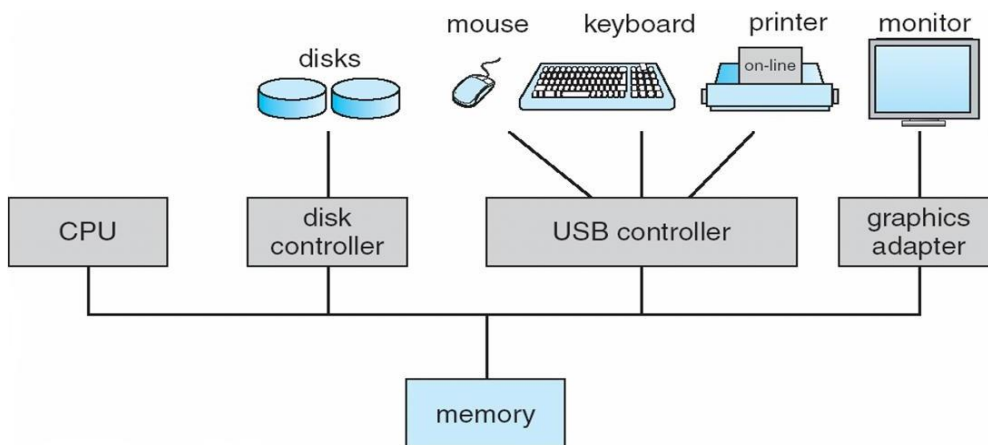
- “The one program running at all times on the computer” is the **kernel**. Everything else is either a system program (ships with the operating system) or an application program

### Computer Startup

- **bootstrap program** is loaded at power-up or reboot
- Typically stored in ROM or EPROM, generally known as **firmware**
- Initializes all aspects of system
- Loads operating system kernel and starts execution

### Computer System Organization

- Computer-system operation
- One or more CPUs, device controllers connect through common bus providing access to shared memory
- Concurrent execution of CPUs and devices competing for memory cycles



### Computer-System Operation

- I/O devices and the CPU can execute concurrently
- Each device controller is in charge of a particular device type
- Each device controller has a local buffer
- CPU moves data from/to main memory to/from local buffers
- I/O is from the device to local buffer of controller
- Device controller informs CPU that it has finished its operation by causing An *interrupt*

### Common Functions of Interrupts

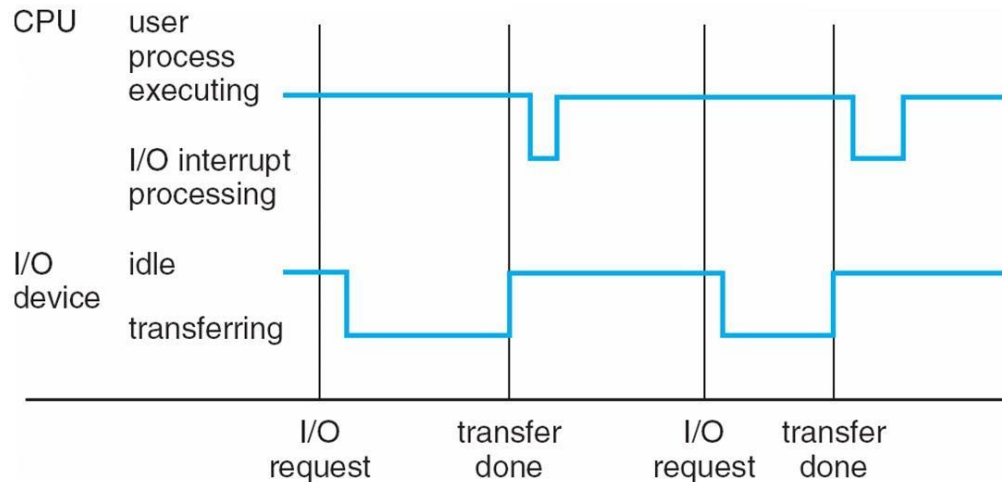
- Interrupt transfers control to the interrupt service routine generally, through the **interrupt vector**, which contains the addresses of all the service routines
- Interrupt architecture must save the address of the interrupted instruction
- Incoming interrupts are *disabled* while another interrupt is being processed to prevent a *lost interrupt* A *trap* is a software-generated interrupt caused either by an error or a user request
- An operating system is **interrupt driven**

### Interrupt Handling

- The operating system preserves the state of the CPU by storing registers and the program counter
- Determines which type of interrupt has occurred:
- **polling**

- **vectored** interrupt system
- Separate segments of code determine what action should be taken for each type of interrupt

### Interrupt Timeline



### I/O Structure

- After I/O starts, control returns to user program only upon I/O completion
- Wait instruction idles the CPU until the next interrupt
- Wait loop (contention for memory access)
- At most one I/O request is outstanding at a time, no simultaneous I/O processing
- After I/O starts, control returns to user program without waiting for I/O completion
- **System call** – request to the operating system to allow user to wait for I/O completion
- **Device-status table** contains entry for each I/O device indicating its type, address, and **state**
- Operating system indexes into I/O device table to determine device status and to modify table entry to include interrupt

### Direct Memory Access Structure

- Used for high-speed I/O devices able to transmit information at close to memory speeds
- Device controller transfers blocks of data from buffer storage directly to main memory without CPU intervention
- Only one interrupt is generated per block, rather than the one interrupt per byte

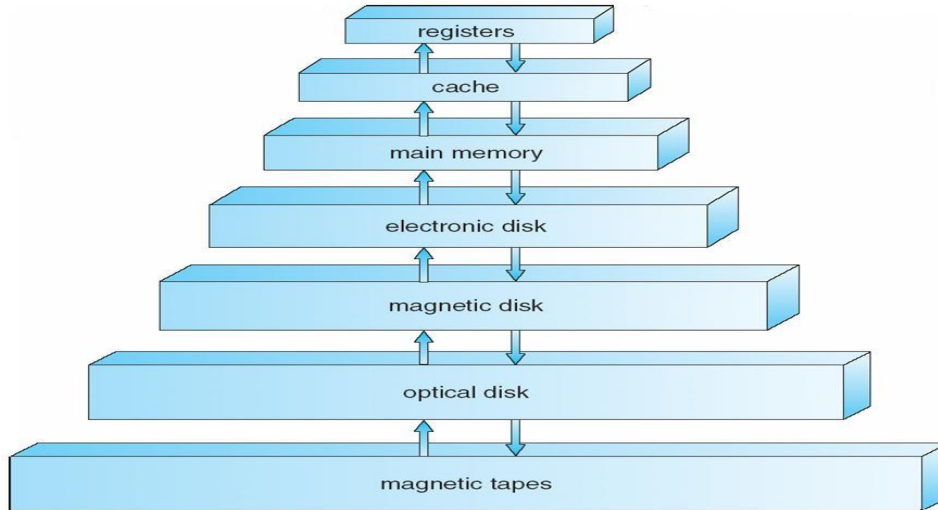
### Storage Structure

- Main memory – only large storage media that the CPU can access directly
- Secondary storage – extension of main memory that provides large nonvolatile storage capacity
- Magnetic disks – rigid metal or glass platters covered with magnetic recording material
- Disk surface is logically divided into **tracks**, which are subdivided into **sectors**
- The **disk controller** determines the logical interaction between the device and the computer

### Storage Hierarchy

- Storage systems organized in hierarchy
  - Speed
  - Cost
  - Volatility
-

**Caching** – copying information into faster storage system; main memory can be viewed as a last *cache* for secondary storage



### **Caching**

- Important principle, performed at many levels in a computer (in hardware, operating system, software)
- Information in use copied from slower to faster storage temporarily
- Faster storage (cache) checked first to determine if information is there
- If it is, information used directly from the cache (fast)
- If not, data copied to cache and used there
- Cache smaller than storage being cached
- Cache management important design problem
- Cache size and replacement policy

### **Computer-System Architecture**

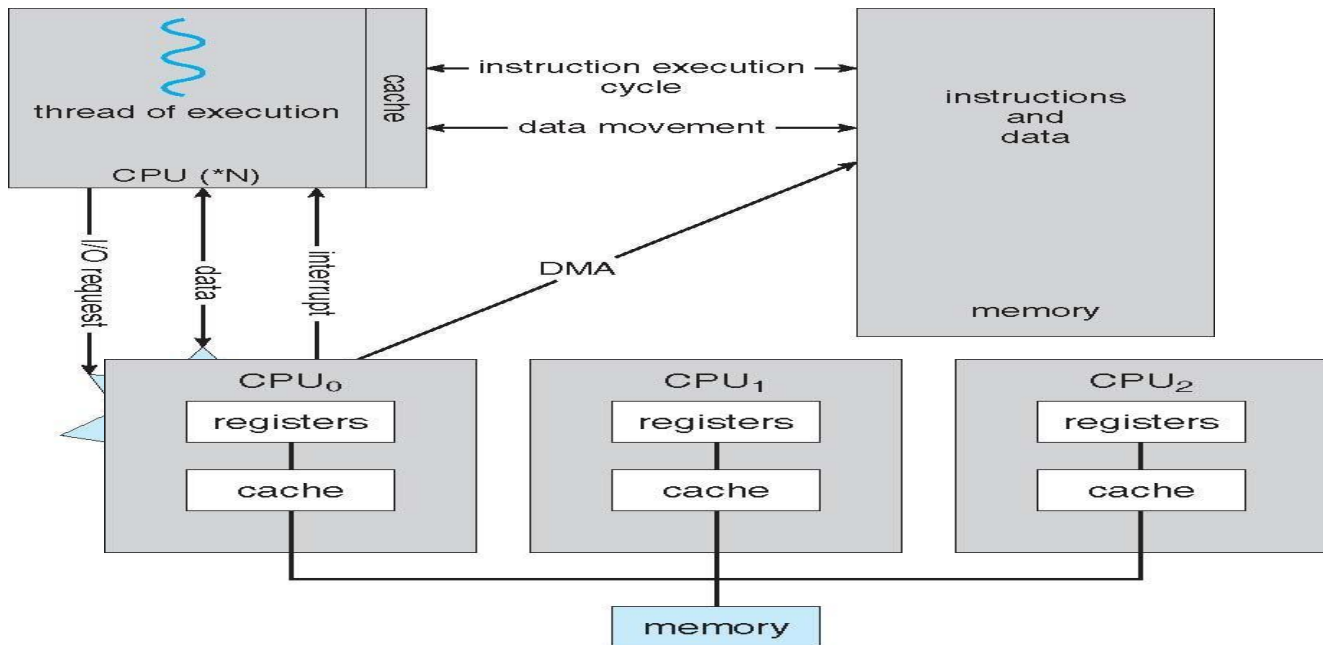
- Most systems use a single general-purpose processor (PDAs through mainframes)
- Most systems have special-purpose processors as well
- Multiprocessors systems growing in use and importance
- Also known as parallel systems, tightly-coupled systems

Advantages include

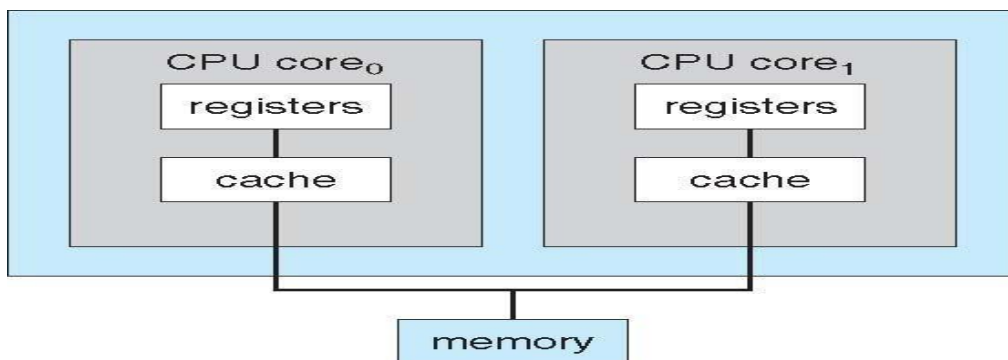
1. Increased throughput
2. Economy of scale
3. Increased reliability – graceful degradation or fault tolerance

Two types

1. Asymmetric Multiprocessing
  2. Symmetric Multiprocessing
-



### A Dual-Core Design



### Clustered Systems

- Like multiprocessor systems, but multiple systems working together
- Usually sharing storage via a storage-area network (SAN)
- Provides a high-availability service which survives failures
  - Asymmetric clustering has one machine in hot-standby mode
  - Symmetric clustering has multiple nodes running applications, monitoring each other
- Some clusters are for high-performance computing (HPC)
  - Applications must be written to use parallelization

### Operating System Structure

- **Multiprogramming** needed for efficiency
- Single user cannot keep CPU and I/O devices busy at all times
- Multiprogramming organizes jobs (code and data) so CPU always has one to Execute

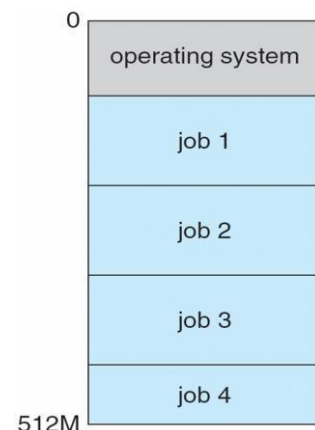
•

A subset of total jobs in system is kept in memory

- One job selected and run via **job scheduling**
- When it has to wait (for I/O for example), OS switches to another job
- **Timesharing (multitasking)** is logical extension in which CPU switches jobs so frequently that users can interact with each job while it is running, creating **interactive** computing
- **Response time** should be < 1 second
- Each user has at least one program executing in memory [**process**
- If several jobs ready to run at the same time [ **CPU scheduling**
- If processes don't fit in memory, **swapping** moves them in and out to run

**Virtual memory** allows execution of processes not completely in memory

### Memory Layout for Multiprogrammed System



### Operating-System Operations

- Interrupt driven by hardware
- Software error or request creates **exception** or **trap**
- Division by zero, request for operating system service
- Other process problems include infinite loop, processes modifying each Other or the operating system
- **Dual-mode** operation allows OS to protect itself and other system components
- **User mode** and **kernel mode**
- **Mode bit** provided by hardware

□ Provides ability to distinguish when system is running user code or kernel code

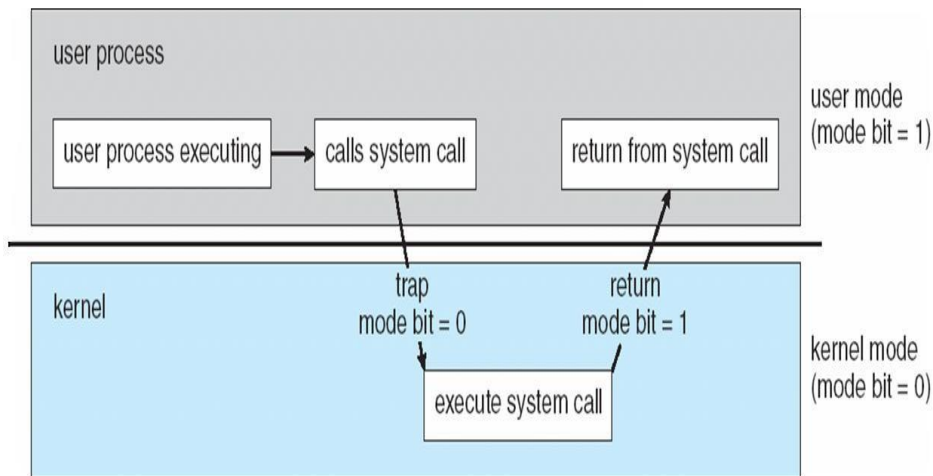
□ Some instructions designated as **privileged**, only executable in kernel mode

□ System call changes mode to kernel, return from call resets it to user

### Transition from User to Kernel Mode

- Timer to prevent infinite loop / process hogging resources
- Set interrupt after specific period
- Operating system decrements counter
- When counter zero generate an interrupt
- Set up before scheduling process to regain control or terminate program that exceeds allotted time





## OPERATING SYSTEM FUNCTIONS

### Process Management

- 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*.
- 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
- Providing mechanisms for process communication
- Providing mechanisms for deadlock handling

### Memory Management

- All data in memory before and after processing
  - All instructions in memory in order to execute
  - Memory management determines what is in memory 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

### 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 dirs
- Mapping files onto secondary storage
- Backup files onto stable (non-volatile) storage media

### Mass-Storage Management

- Usually disks used to store data that does not fit in main memory or data that must be kept for a “long” period of time
- Proper management is of central importance
- Entire speed of computer operation hinges on disk subsystem and its algorithms
- **MASS STORAGE activities**
- Free-space management
- Storage allocation
- Disk scheduling
- Some storage need not be fast
- Tertiary storage includes optical storage, magnetic tape
- Still must be managed
- Varies between WORM (write-once, read-many-times) and RW (read-write)

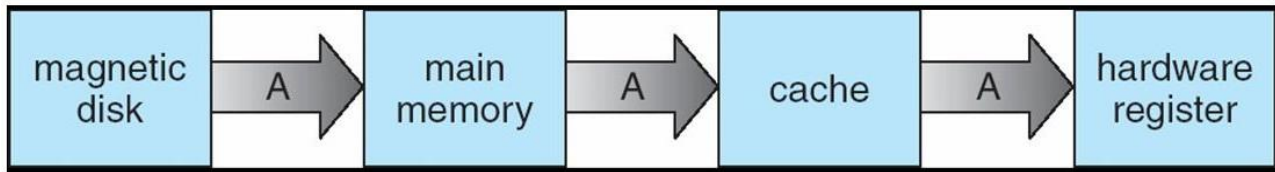
### Performance of Various Levels of Storage

| Level                     | 1                                       | 2                             | 3                | 4                |
|---------------------------|---|-------------------------------|------------------|------------------|
| Name                      | registers                               | cache                         | main memory      | disk storage     |
| Typical size              | < 1 KB                                  | > 16 MB                       | > 16 GB          | > 100 GB         |
| Implementation technology | custom memory with multiple ports, CMOS | on-chip or off-chip CMOS SRAM | CMOS DRAM        | magnetic disk    |
| Access time (ns)          | 0.25 – 0.5                              | 0.5 – 25                      | 80 – 250         | 5,000.000        |
| Bandwidth (MB/sec)        | 20,000 – 100,000                        | 5000 – 10,000                 | 1000 – 5000      | 20 – 150         |
| Managed by                | compiler                                | hardware                      | operating system | operating system |
| Backed by                 | cache                                   | main memory                   | disk             | CD or tape       |

Der

## Migration of Integer A from Disk to Register

- Multitasking environments must be careful to use most recent value, no matter where it is stored in the storage hierarchy



- Multiprocessor environment must provide cache coherency in hardware such that all CPUs have the most recent value in their cache
- Distributed environment situation even more complex
- Several copies of a datum can exist

## 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), spooling (the overlapping of output of one job with input of other jobs)
- General device-driver interface
- Drivers for specific hardware devices

## Protection and Security

**Protection** – any mechanism for controlling access of processes or users to resources defined by the OS

**Security** – defense of the system against internal and external attacks

- Huge range, including denial-of-service, worms, viruses, identity theft, theft of service
- Systems generally first distinguish among users, to determine who can do what
- User identities (**user IDs**, security IDs) include name and associated number, one per user
- User ID then associated with all files, processes of that user to determine access control
- Group identifier (**group ID**) allows set of users to be defined and controls managed, then also associated with each process, file
- **Privilege escalation** allows user to change to effective ID with more rights

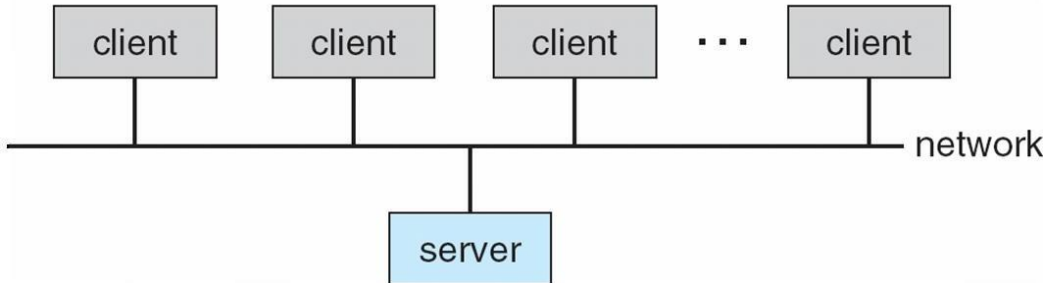
## DISTRIBUTED SYSTEMS

### Computing Environments

#### Traditional computer

- Blurring over time
- Office environment
  - PCs connected to a network, terminals attached to mainframe or minicomputers providing batch and timesharing
  - Now portals allowing networked and remote systems access to same resources
- Home networks
  - Used to be single system, then modems
  - Now firewalled, networked
- Client-Server Computing

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



### Peer-to-Peer Computing

- 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*

### Web-Based Computing

- Web has become ubiquitous
- PCs most prevalent devices
- More devices becoming networked to allow web access
- New category of devices to manage web traffic among similar servers: **load balancers**
- Use of operating systems like Windows 95, client-side, have evolved into Linux and Windows XP, which can be clients and servers

### Open-Source Operating Systems

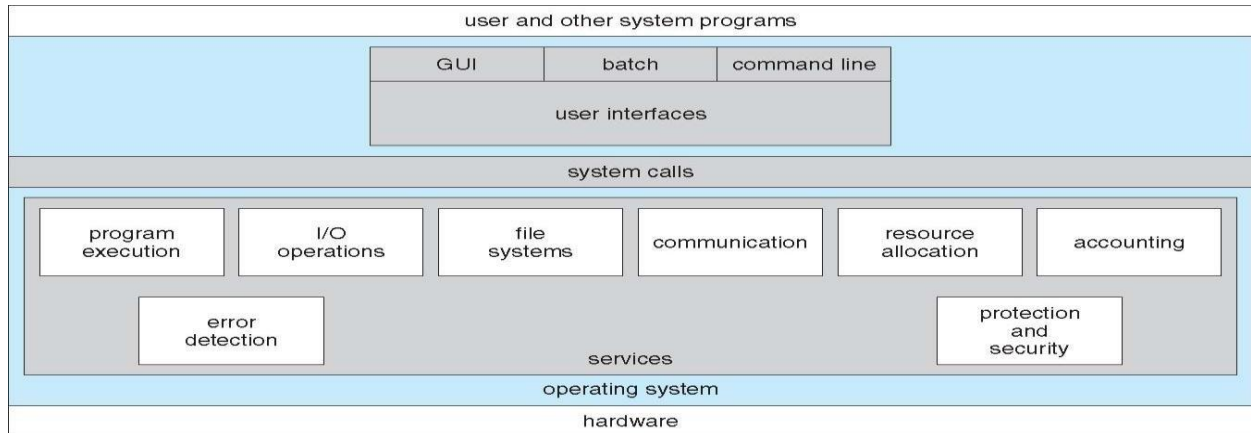
- Operating systems made available in source-code format rather than just binary closed-source
- Counter to the copy protection and Digital Rights Management (DRM) movement
- Started by Free Software Foundation (FSF), which has “copyleft” GNU Public License (GPL)
- Examples include GNU/Linux, BSD UNIX (including core of Mac OS X), and Sun Solaris

### Operating System Services

- 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

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

## A View of Operating System Services



## Operating System Services

- One set of operating-system services provides functions that are helpful to the user
  - 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.
- 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 - Some (such as CPU cycles, main memory, and file storage) may have special allocation code, others (such as I/O devices) may have general request and release code.
- **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.
- If a system is to be protected and secure, precautions must be instituted throughout it. A chain is only as strong as its weakest link.

## User Operating System Interface - CLI

- Command Line Interface (CLI) or command interpreter allows direct command entry

- Sometimes implemented in kernel, sometimes by systems program
- Sometimes multiple flavors implemented – shells
- Primarily fetches a command from user and executes it

- Sometimes commands built-in, sometimes just names of programs
- If the latter, adding new features doesn't require shell modification

## User Operating System Interface - GUI

- User-friendly desktop metaphor interface
- Usually mouse, keyboard, and monitor
- Icons represent files, programs, actions, etc
- Various mouse buttons over objects in the interface cause various actions (provide information, options, execute function, open directory (known as a folder))
- Invented at Xerox PARC
- Many systems now include both CLI and GUI interfaces
- Microsoft Windows is GUI with CLI “command” shell
- Apple Mac OS X as “Aqua” GUI interface with UNIX kernel underneath and shells available
- Solaris is CLI with optional GUI interfaces (Java Desktop, KDE)

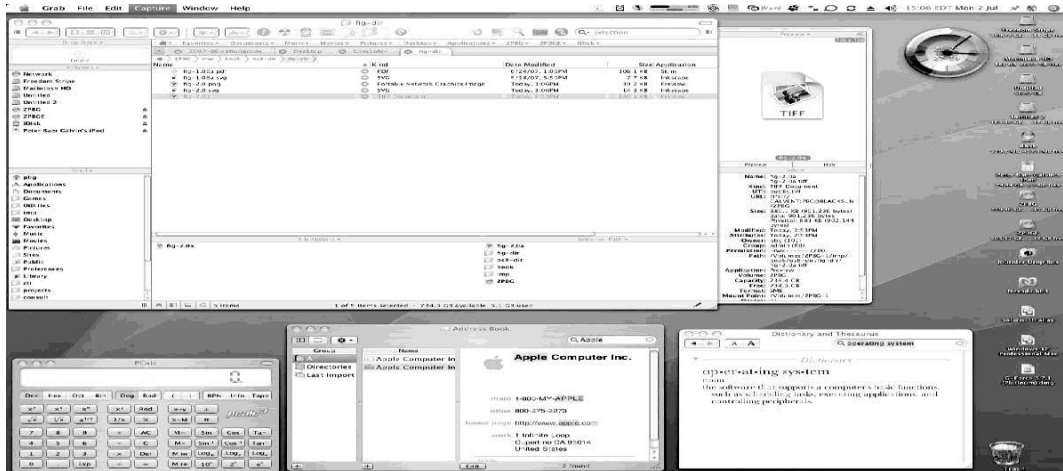
## Bourne Shell Command Interpreter

```

Terminal
File Edit View Terminal Tabs Help
fd0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0 0
sd0 0.0 0.2 0.0 0.2 0.0 0.0 0.0 0.4 0 0
sd1 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0 0
extended device statistics
device r/s w/s kr/s kw/s wait actv svc_t %w %b
fd0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0 0
sd0 0.6 0.0 38.4 0.0 0.0 0.0 8.2 0 0
sd1 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0 0
(root@pbg-nv64-vm)-(11/pts)-(00:53 15-Jun-2007)-(global)
- (/var/tmp/system-contents/scripts)# swap -sh
total: 1.1G allocated + 190M reserved = 1.3G used, 1.6G available
(root@pbg-nv64-vm)-(12/pts)-(00:53 15-Jun-2007)-(global)
- (/var/tmp/system-contents/scripts)# uptime
12:53am up 9 min(s), 3 users, load average: 33.29, 67.68, 36.81
(root@pbg-nv64-vm)-(13/pts)-(00:53 15-Jun-2007)-(global)
- (/var/tmp/system-contents/scripts)# w
 4:07pm up 17 day(s), 15:24, 3 users, load average: 0.09, 0.11, 8.66
User  tty      login@ idle  JCPU  PCPU  what
root  console  15Jun0718days  1      /usr/bin/ssh-agent -- /usr/bi
n/d
root  pts/3    15Jun07      18     4 w
root  pts/4    15Jun0718days      w
(root@pbg-nv64-vm)-(14/pts)-(16:07 02-Jul-2007)-(global)
- (/var/tmp/system-contents/scripts)#

```

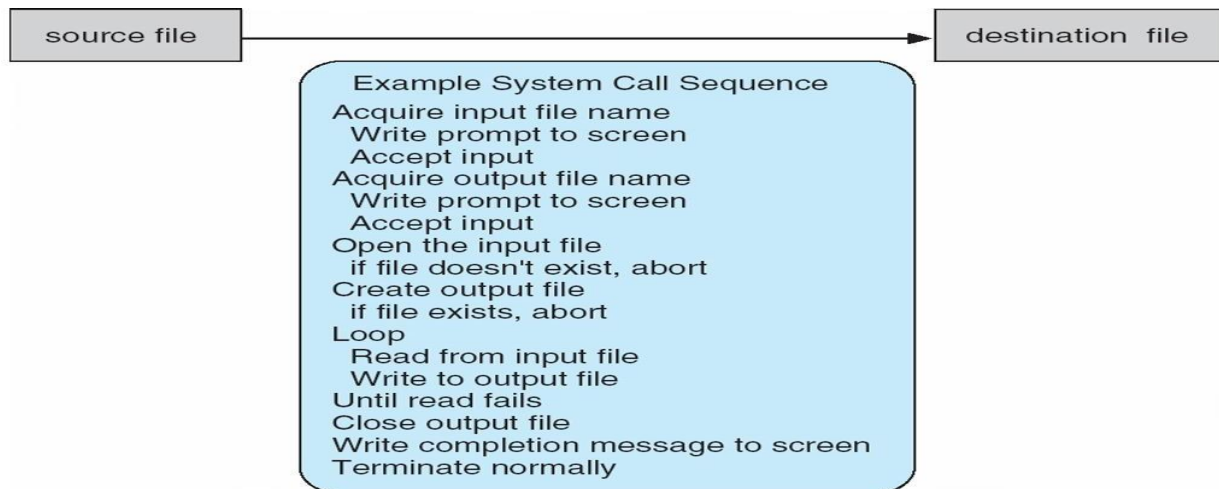
## The Mac OS X GUI



## System Calls

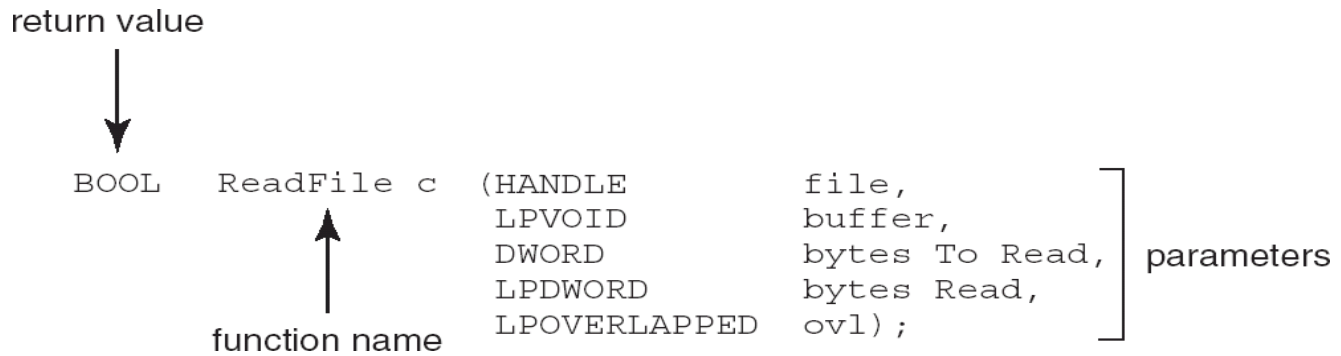
- Programming interface to the services provided by the OS
- Typically written in a high-level language (C or C++)
- Mostly accessed by programs via a high-level Application Program Interface (API) rather than direct system call use. Three most common APIs are Win32 API for Windows, POSIX API for POSIX-based systems (including virtually all versions of UNIX, Linux, and Mac OS X), and Java API for the Java virtual machine (JVM)
- Why use APIs rather than system calls? (Note that the system-call names used throughout this text are generic)

## Example of System Calls



## Example of Standard API

Consider the ReadFile() function in the Win32 API—a function for reading from a file



A description of the parameters passed to ReadFile()

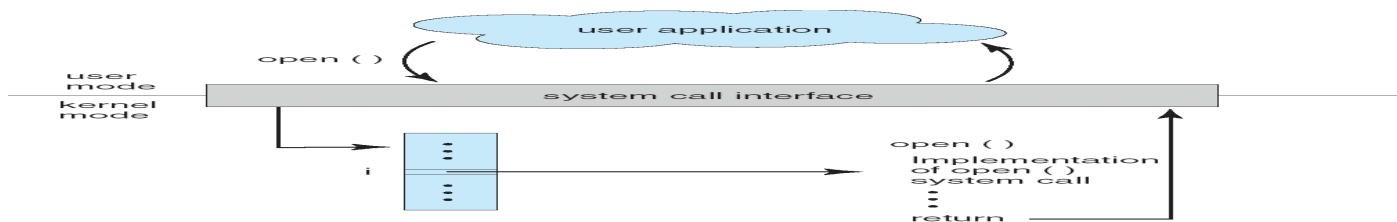
- HANDLE file—the file to be read
- LPVOID buffer—a buffer where the data will be read into and written from
- DWORD bytesToRead—the number of bytes to be read into the buffer
- LPDWORD bytesRead—the number of bytes read during the last read
- LPOVERLAPPED ovl—indicates if overlapped I/O is being used

## System Call Implementation

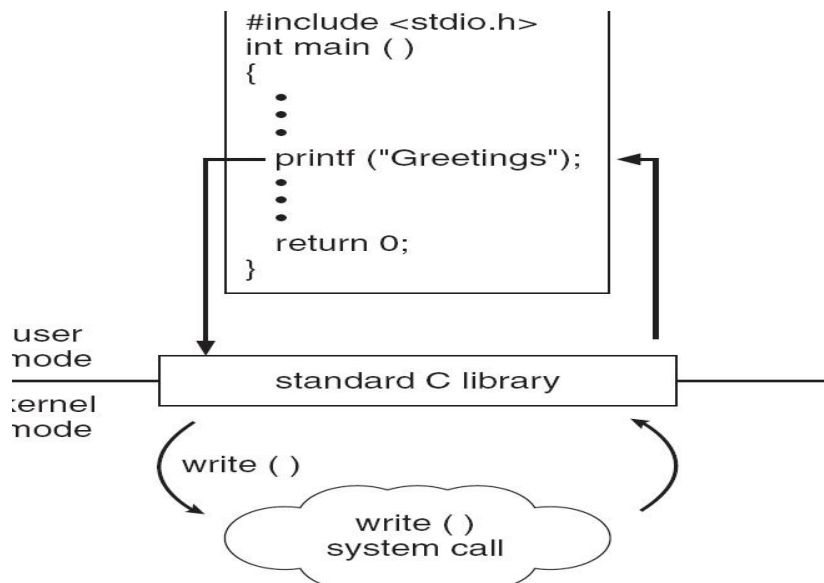
- Typically, a number associated with each system call
- System-call interface maintains a table indexed according to these Numbers
- The system call interface invokes intended system call in OS kernel and returns status of the system call and any return values
- The caller need know nothing about how the system call is implemented
- Just needs to obey API and understand what OS will do as a result call
- Most details of OS interface hidden from programmer by API

Managed by run-time support library (set of functions built into libraries included with compiler)

## API – System Call – OS Relationship



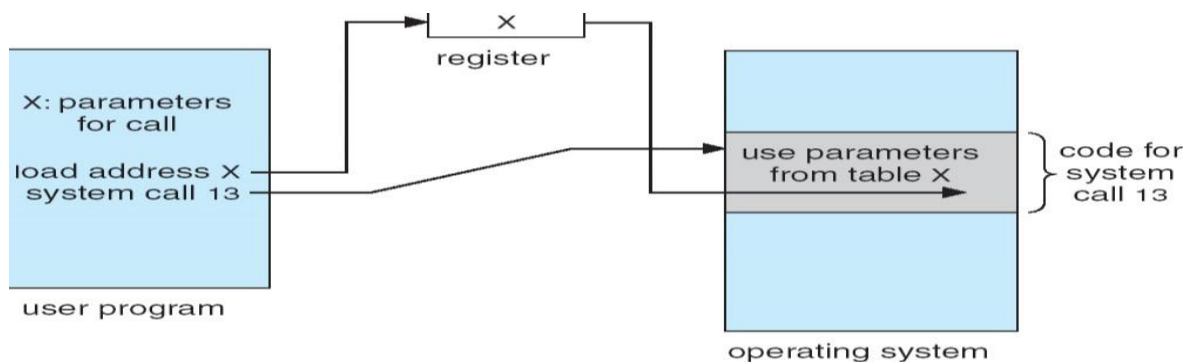




### System Call Parameter Passing

- Often, more information is required than simply identity of desired system call
- Exact type and amount of information vary according to OS and call
- Three general methods used to pass parameters to the OS
- Simplest: pass the parameters in *registers*
  - In some cases, may be more parameters than registers
- Parameters stored in a *block*, or table, in memory, and address of block passed as a parameter in a register
  - This approach taken by Linux and Solaris
- Parameters placed, or *pushed*, onto the *stack* by the program and *popped* off the stack by the operating system
- Block and stack methods do not limit the number or length of parameters being passed

### Parameter Passing via Table



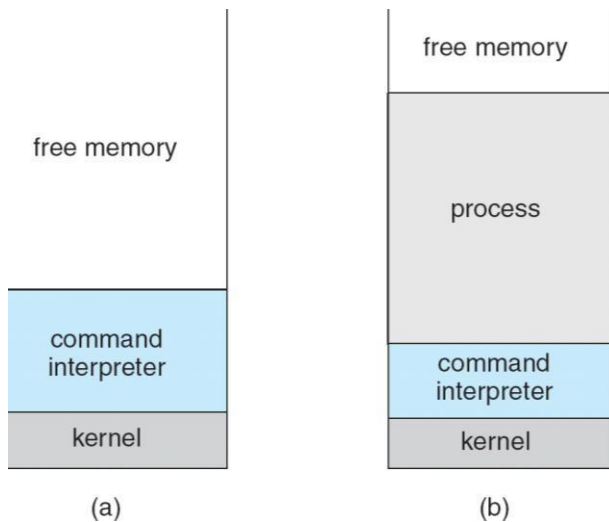
## Types of System Calls

- Process control
- File management
- Device management
- Information maintenance
- Communications
- Protection

## Examples of Windows and Unix System Calls

|                                | Windows   | Unix                                   |
|--------------------------------|---|--|
| <b>Process Control</b>         | CreateProcess()<br>ExitProcess()<br>WaitForSingleObject()                           | fork()<br>exit()<br>wait()             |
| <b>File Manipulation</b>       | CreateFile()<br>ReadFile()<br>WriteFile()<br>CloseHandle()                          | open()<br>read()<br>write()<br>close() |
| <b>Device Manipulation</b>     | SetConsoleMode()<br>ReadConsole()<br>WriteConsole()                                 | ioctl()<br>read()<br>write()           |
| <b>Information Maintenance</b> | GetCurrentProcessID()<br>SetTimer()<br>Sleep()                                      | getpid()<br>alarm()<br>sleep()         |
| <b>Communication</b>           | CreatePipe()<br>CreateFileMapping()<br>MapViewOfFile()                              | pipe()<br>shmget()<br>mmap()           |
| <b>Protection</b>              | SetFileSecurity()<br>InitializeSecurityDescriptor()<br>SetSecurityDescriptorGroup() | chmod()<br>umask()<br>chown()          |

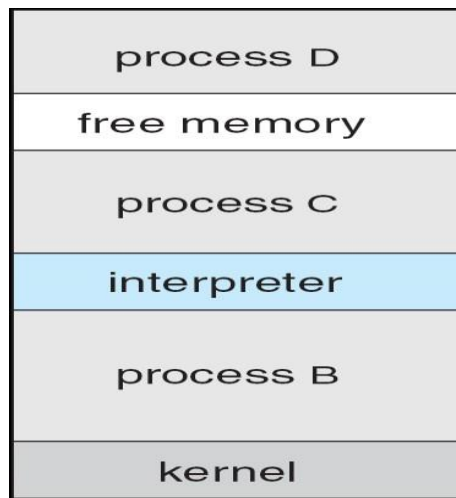
## MS-DOS execution



(a) At system startup

(b) running a program

## FreeBSD Running Multiple Programs



### System Programs

System programs provide a convenient environment for program development and execution. They can be divided into:

- File manipulation
- Status information
- File modification
- Programming language support
- Program loading and execution
- Communications
- Application programs

Most users' view of the operation system is defined by system programs, not the actual system calls

- Provide a convenient environment for program development and execution
- Some of them are simply user interfaces to system calls; others are considerably more complex
- File management - Create, delete, copy, rename, print, dump, list, and generally manipulate files and directories
- Status information
- Some ask the system for info - date, time, amount of available memory, disk space, number of users
- Others provide detailed performance, logging, and debugging information
- Typically, these programs format and print the output to the terminal or other output devices
- Some systems implement a registry - used to store and retrieve configuration information

File modification

- Text editors to create and modify files
  - Special commands to search contents of files or perform transformations of the text
  - Programming-language support - Compilers, assemblers, debuggers and interpreters sometimes provided
  - Program loading and execution- Absolute loaders, relocatable loaders, linkage editors, and overlay-loaders, debugging systems for higher-level and machine language
-

- Communications - Provide the mechanism for creating virtual connections among processes, users, and computer systems
- Allow users to send messages to one another's screens, browse web pages, send electronic-mail messages, log in remotely, transfer files from one machine to another

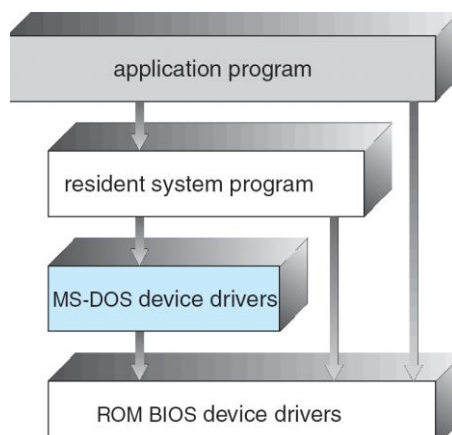
### Operating System Design and Implementation

- Design and Implementation of OS not “solvable”, but some approaches have proven successful
- Internal structure of different Operating Systems can vary widely
- Start by defining goals and specifications
- Affected by choice of hardware, type of system
- *User goals and System goals*
- User goals – operating system should be convenient to use, easy to learn, reliable, safe, and fast
- System goals – operating system should be easy to design, implement, and maintain, as well as flexible, reliable, error-free, and efficient
- Important principle to separate
- **Policy:** What will be done?  
**Mechanism:** How to do it?
- Mechanisms determine how to do something, policies decide what will be done
- The separation of policy from mechanism is a very important principle, it allows maximum flexibility if policy decisions are to be changed later

### Simple Structure

- MS-DOS – written to provide the most functionality in the least space
- Not divided into modules
- Although MS-DOS has some structure, its interfaces and levels of Functionality are not well separated

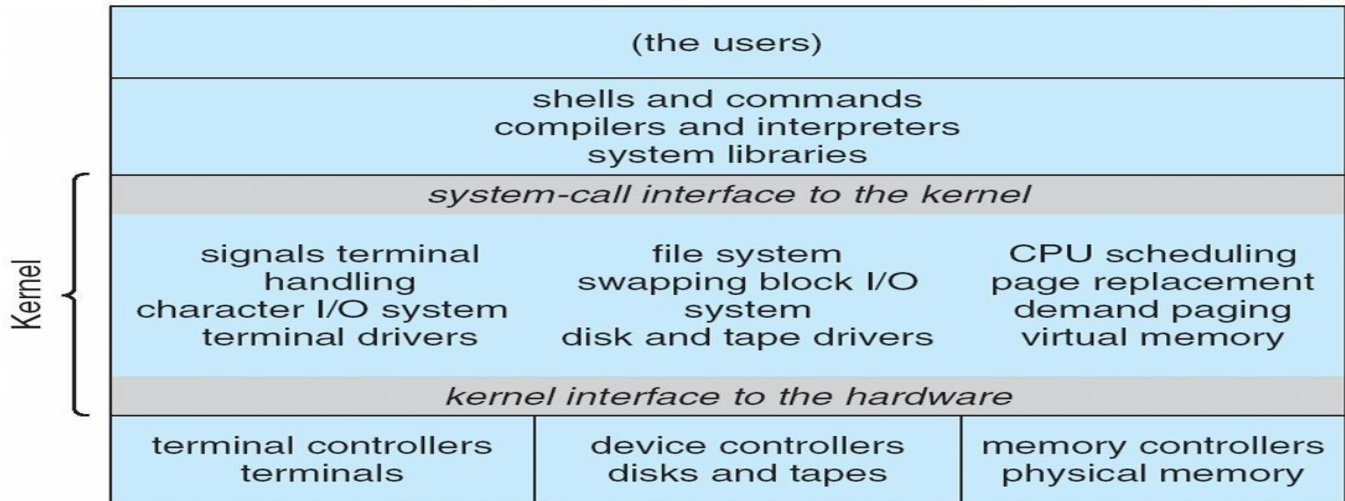
### MS-DOS Layer Structure



### Layered Approach

- The operating system is divided into a number of layers (levels), each built on top of lower layers. The bottom layer (layer 0), is the hardware; the highest (layer N) is the user interface.
  - With modularity, layers are selected such that each uses functions (operations) and services of only lower-level layers
-

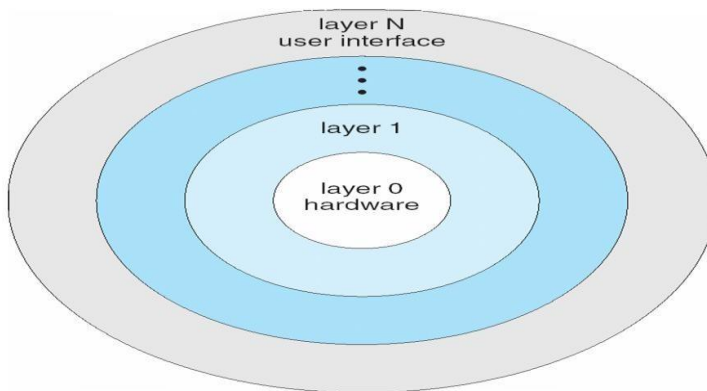
## Traditional UNIX System Structure



## UNIX

- UNIX – limited by hardware functionality, the original UNIX operating system had limited structuring. The UNIX OS consists of two separable parts
- Systems programs
- The kernel
  - Consists of everything below the system-call interface and above the physical hardware
  - Provides the file system, CPU scheduling, memory management, and other operating-system functions; a large number of functions for one level

## Layered Operating System

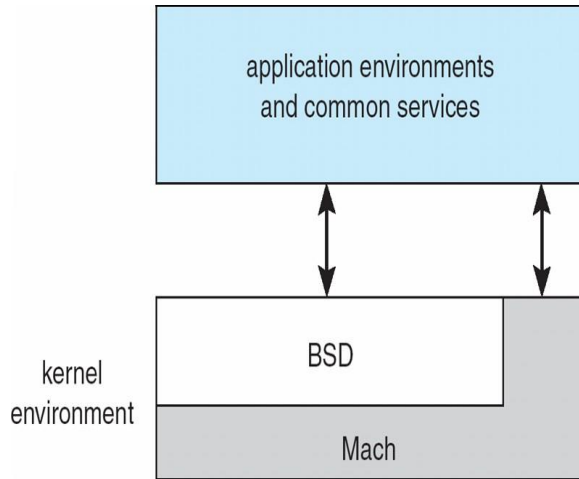


## Micro kernel System Structure

- Moves as much from the kernel into “user” space
- Communication takes place between user modules using message passing
- Benefits:
  - Easier to extend a microkernel
  - Easier to port the operating system to new architectures
  - More reliable (less code is running in kernel mode)

- More secure
- Detriments:
- Performance overhead of user space to kernel space communication

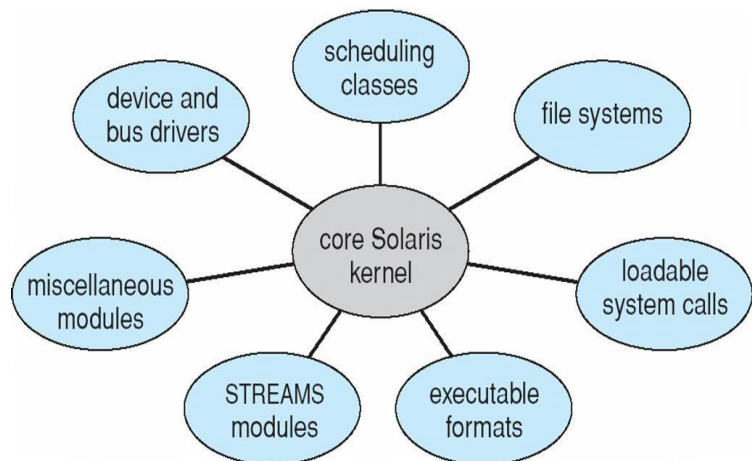
### Mac OS X Structure



### Modules

- Most modern operating systems implement kernel modules
- Uses object-oriented approach
- Each core component is separate
- Each talks to the others over known interfaces
- Each is loadable as needed within the kernel
- Overall, similar to layers but with more flexible

### Solaris Modular Approach



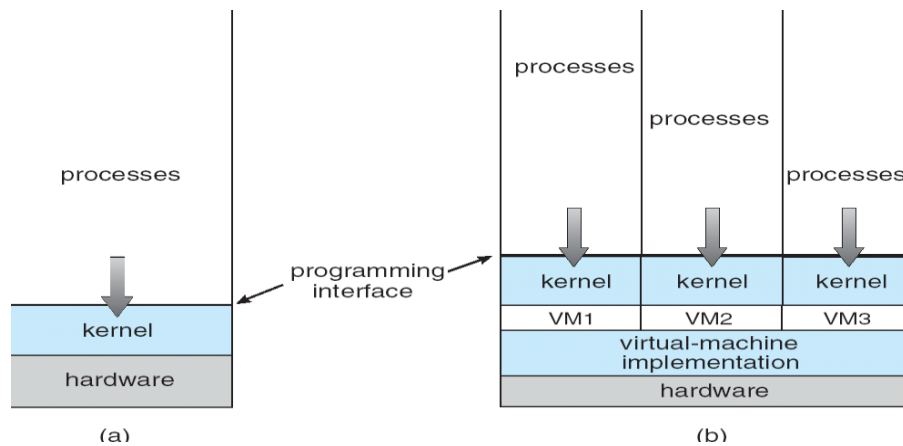
### Virtual Machines

- A virtual machine takes the layered approach to its logical conclusion. It treats hardware and the operating system kernel as though they were all hardware

- A virtual machine provides an interface *identical* to the underlying bare hardware
- The operating system host creates the illusion that a process has its own processor and (virtual memory)
- Each guest provided with a (virtual) copy of underlying computer

### Virtual Machines History and Benefits

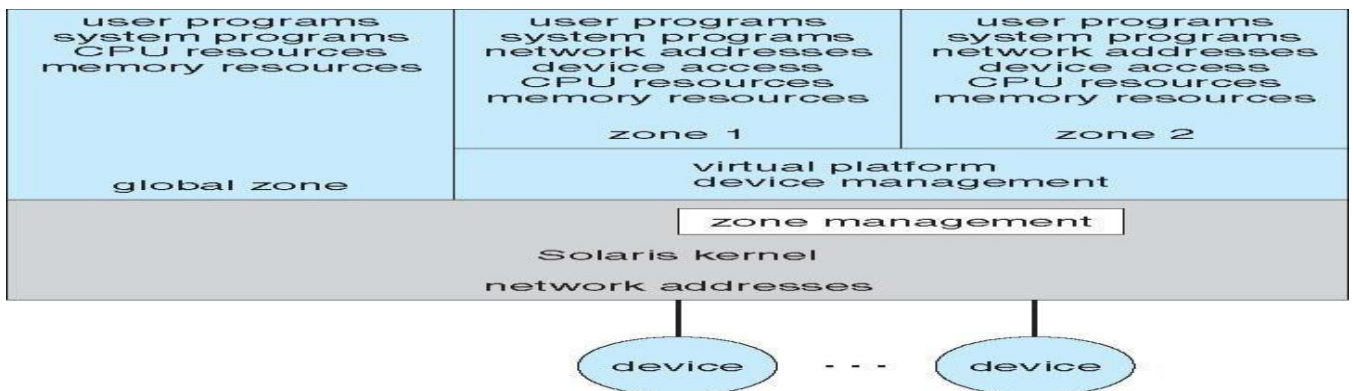
- First appeared commercially in IBM mainframes in 1972
- Fundamentally, multiple execution environments (different operating systems) can share the same hardware
- Protect from each other
- Some sharing of file can be permitted, controlled
- Commutate with each other, other physical systems via networking
- Useful for development, testing
- Consolidation of many low-resource use systems onto fewer busier systems
- “Open Virtual Machine Format”, standard format of virtual machines, allows a VM to run within many different virtual machine (host) platforms



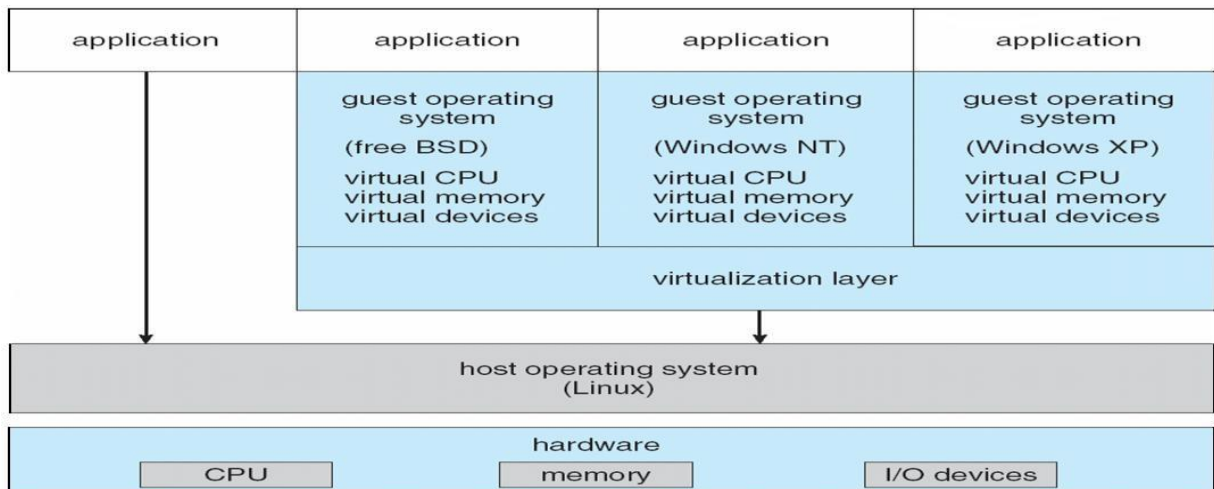
### Para-virtualization

- Presents guest with system similar but not identical to hardware
- Guest must be modified to run on paravirtualized hardware
- Guest can be an OS, or in the case of Solaris 10 applications running in containers

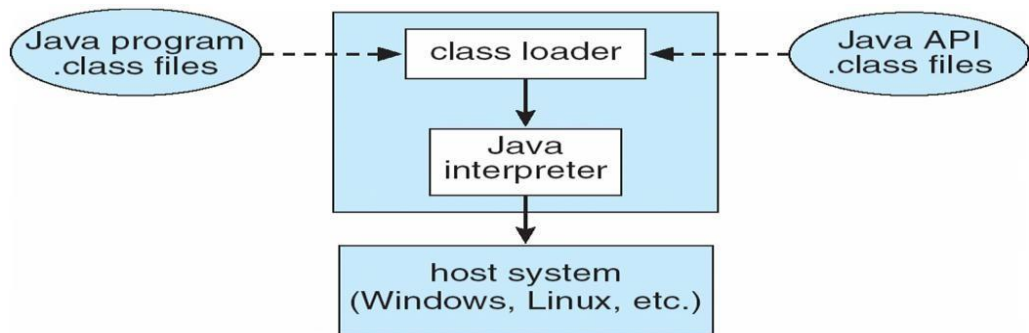
### Solaris 10 with Two Containers



## VMware Architecture



## The Java Virtual Machine



## Operating-System Debugging

- Debugging is finding and fixing errors, or bugs
- OSes generate log files containing error information
- Failure of an application can generate core dump file capturing memory of the process
- Operating system failure can generate crash dump file containing kernel memory
- Beyond crashes, performance tuning can optimize system performance
- Kernighan's Law: "Debugging is twice as hard as writing the code in the rst place. Therefore, if you write the code as cleverly as possible, you are, by definition, not smart enough to debug it."
- DTrace tool in Solaris, FreeBSD, Mac OS X allows live instrumentation on production systems
- Probes fire when code is executed, capturing state data and sending it to consumers of those probes



## Solaris 10 dtrace Following System Call

### Operating System Generation

- Operating systems are designed to run on any of a class of machines; the system must be configured for each specific computer site
- SYSGEN program obtains information concerning the specific configuration of the hardware system
- *Booting* – starting a computer by loading the kernel
- *Bootstrap program* – code stored in ROM that is able to locate the kernel, load it into memory, and start its execution

### System Boot

- Operating system must be made available to hardware so hardware can start it
  - Small piece of code – **bootstrap loader**, locates the kernel, loads it into memory, and starts it
  - Sometimes two-step process where **boot block** at fixed location loads bootstrap loader
  - When power initialized on system, execution starts at a fixed memory location Firmware used to hold initial boot code
-

# UNIT -2



# PROCESS MANAGEMENT

## Process Concept

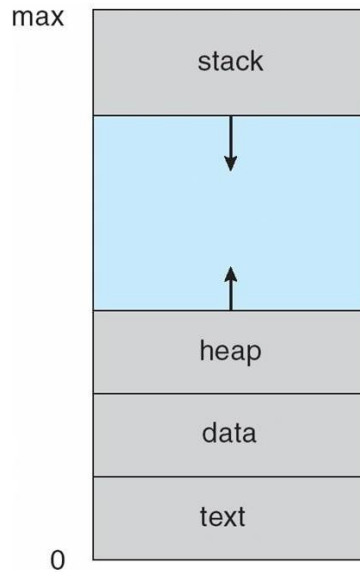
- An operating system executes a variety of programs:
- Batch system – jobs
- Time-shared systems – user programs or tasks
- Textbook uses the terms *job* and *process* almost interchangeably

Process – a program in execution; process execution must progress in sequential fashion

A process includes:

- program counter
- stack
- data section

## Process in Memory

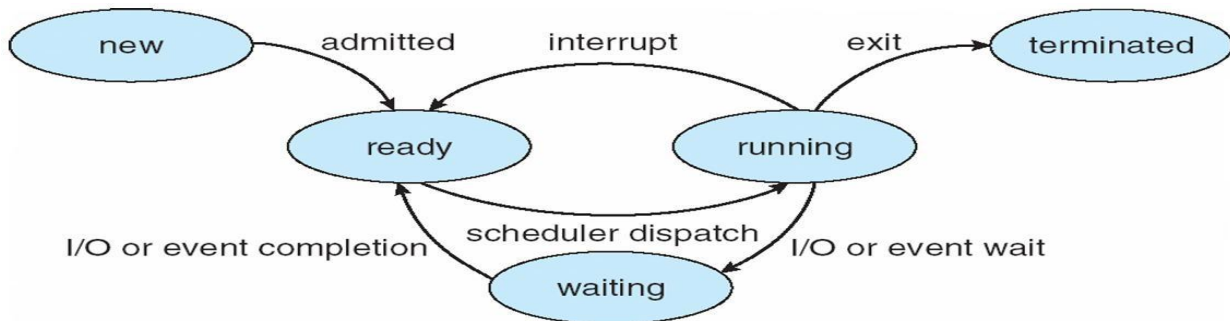


## Process State

As a process executes, it changes *state*

- **new**: The process is being created
- **running**: Instructions are being executed
- **waiting**: The process is waiting for some event to occur
- **ready**: The process is waiting to be assigned to a processor
- **terminated**: The process has finished execution

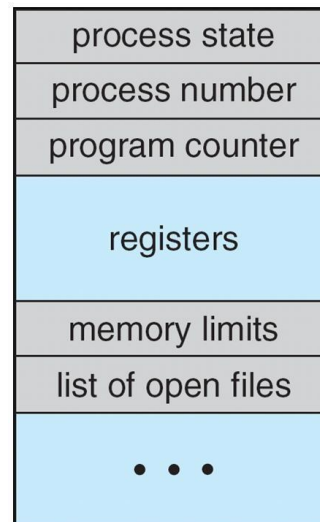
## Diagram of Process State



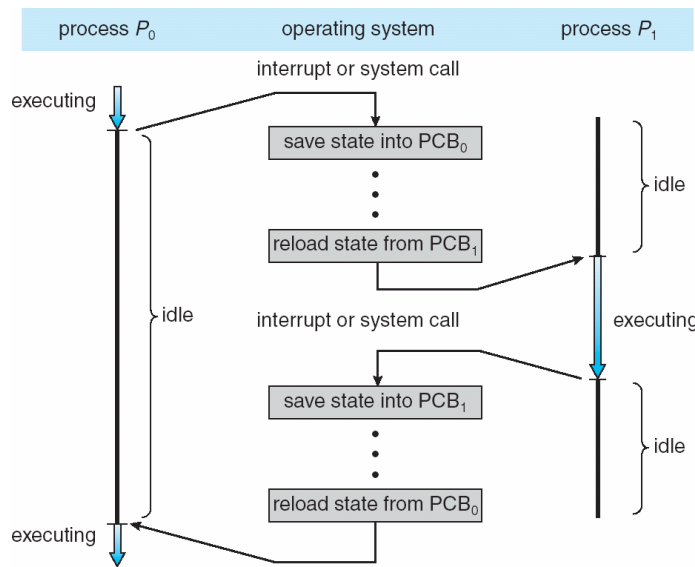
## Process Control Block (PCB)

Information associated with each process

- Process state
- Program counter
- CPU registers
- CPU scheduling information
- Memory-management information
- Accounting information
- I/O status information



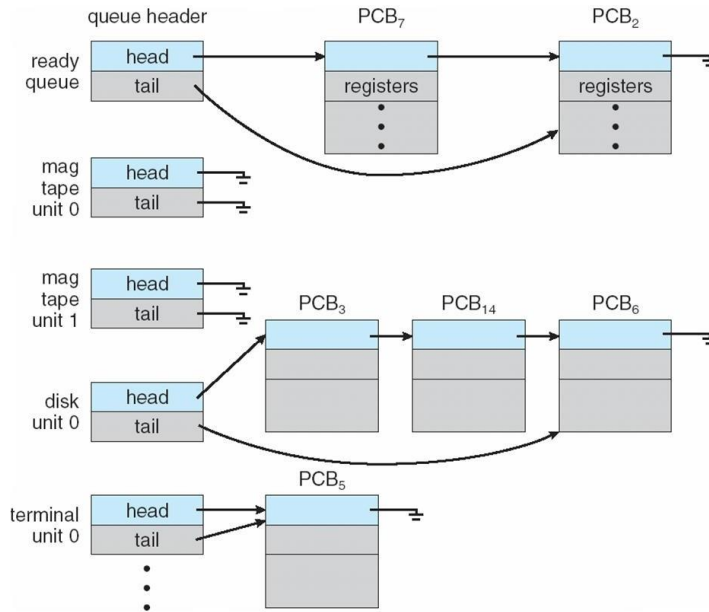
## CPU Switch From Process to Process



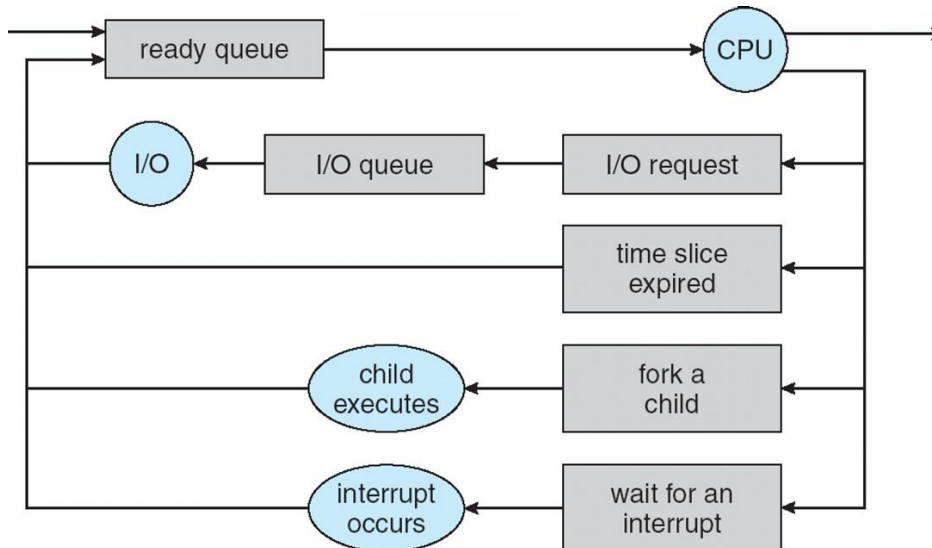
## Process Scheduling Queues

- **Job queue** – set of all processes in the system
- **Ready queue** – set of all processes residing in main memory, ready and waiting to execute
- **Device queues** – set of processes waiting for an I/O device
- Processes migrate among the various queues

## Ready Queue and Various I/O Device Queues



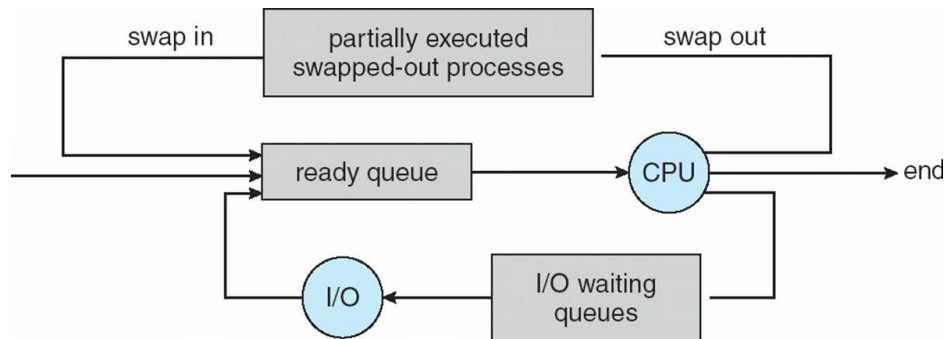
## Representation of Process Scheduling



## Schedulers

- **Long-term scheduler** (or job scheduler) – selects which processes should be brought into the ready queue
- **Short-term scheduler** (or CPU scheduler) – selects which process should be executed next and allocates CPU

## Addition of Medium Term Scheduling



- Short-term scheduler is invoked very frequently (milliseconds)  $\mathcal{P}$  (must be fast)
- Long-term scheduler is invoked very infrequently (seconds, minutes)  $\mathcal{P}$  (may be slow)
- The long-term scheduler controls the *degree of multiprogramming*
- Processes can be described as either:
- **I/O-bound process** – spends more time doing I/O than computations, many short CPU bursts
- **CPU-bound process** – spends more time doing computations; few very long CPU bursts

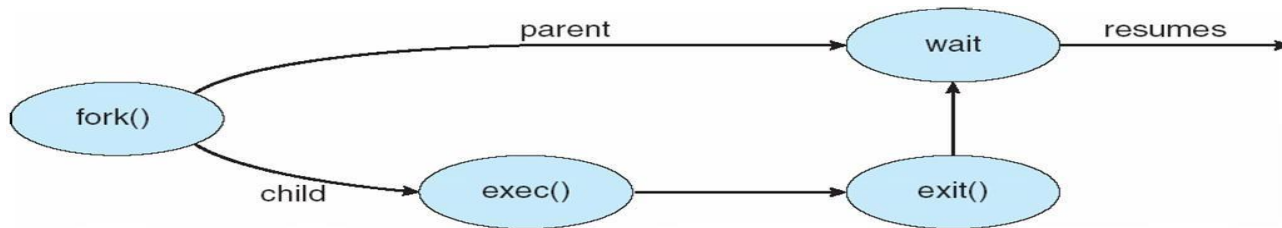
### Context Switch

- When CPU switches to another process, the system must save the state of the old process and load the saved state for the new process via a context switch
- Context of a process represented in the PCB
- Context-switch time is overhead; the system does no useful work while switching
- Time dependent on hardware support

### Process Creation

- **Parent** process create **children** processes, which, in turn create other processes, forming a tree of processes
  - Generally, process identified and managed via a **process identifier (pid)**
  - Resource sharing
  - Parent and children share all resources
  - Children share subset of parent's resources
  - Parent and child share no resources
  - Execution
  - Parent and children execute concurrently
  - Parent waits until children terminate
  - Address space
  - Child duplicate of parent
  - Child has a program loaded into it
  - UNIX examples
  - **fork** system call creates new process
  - **exec** system call used after a **fork** to replace the process' memory space with a new program
-

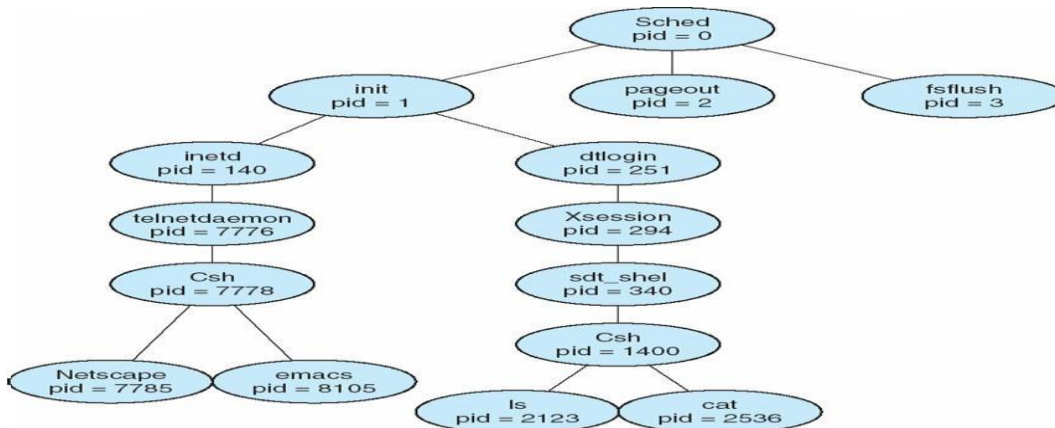
## Process Creation



## C Program Forking Separate Process

```
int main()
{
pid_t pid;
    /* fork another process */
    pid = fork();
    if (pid < 0) { /* error occurred */
        fprintf(stderr, "Fork Failed");
        exit(-1);
    }
    else if (pid == 0) { /* child process */
        execlp("/bin/ls", "ls", NULL);
    }
    else { /* parent process */
        /* parent will wait for the child to complete */
        wait (NULL);
        printf ("Child Complete");
        exit(0);
    }
}
```

## A tree of processes on a typical Solaris



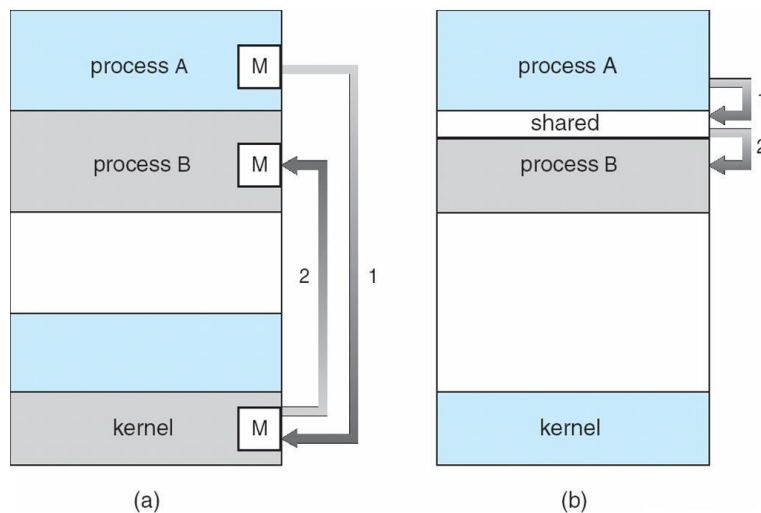
## Process Termination

- Process executes last statement and asks the operating system to delete it (**exit**)
- Output data from child to parent (via **wait**)
- Process' resources are deallocated by operating system
- Parent may terminate execution of children processes (**abort**)
- Child has exceeded allocated resources
- Task assigned to child is no longer required
- If parent is exiting Some operating system do not allow child to continue if its parent terminates  
All children terminated - **cascading termination**

## Interprocess Communication

- Processes within a system may be **independent** or **cooperating**
- Cooperating process can affect or be affected by other processes, including sharing data
- Reasons for cooperating processes:
  - Information sharing
  - Computation speedup
  - Modularity
  - Convenience
- Cooperating processes need **interprocess communication (IPC)**
- Two models of IPC
  - Shared memory
  - Message passing

## Communications Models



## Cooperating Processes

- **Independent** process cannot affect or be affected by the execution of another process
- **Cooperating** process can affect or be affected by the execution of another process

## Advantages of process cooperation

- Information sharing
  - Computation speed-up
  - Modularity
  - Convenience
-



## Producer-Consumer Problem

- Paradigm for cooperating processes, *producer* process produces information that is consumed by a *consumer* process
- *unbounded-buffer* places no practical limit on the size of the buffer
- *bounded-buffer* assumes that there is a fixed buffer size

## Direct Communication

- Processes must name each other explicitly:
- **send**(*P*, *message*) – send a message to process *P*
- **receive**(*Q*, *message*) – receive a message from process *Q*
- Properties of communication link
- Links are established automatically
- A link is associated with exactly one pair of communicating processes
- Between each pair there exists exactly one link
- The link may be unidirectional, but is usually bi-directional

## Indirect Communication

- Messages are directed and received from mailboxes (also referred to as ports)
- Each mailbox has a unique id
- Processes can communicate only if they share a mailbox
- Properties of communication link
- Link established only if processes share a common mailbox
- A link may be associated with many processes
- Each pair of processes may share several communication links
- Link may be unidirectional or bi-directional
- Operations
- create a new mailbox
- send and receive messages through mailbox
- destroy a mailbox
- Primitives are defined as:
- **send**(*A*, *message*) – send a message to mailbox *A*
- **receive**(*A*, *message*) – receive a message from mailbox *A*
- Mailbox sharing
- *P1*, *P2*, and *P3* share mailbox *A*
- *P1*, sends; *P2* and *P3* receive
- Who gets the message?
- Solutions
- Allow a link to be associated with at most two processes
- Allow only one process at a time to execute a receive operation
- Allow the system to select arbitrarily the receiver. Sender is notified who the receiver was.

## Synchronization

- Message passing may be either blocking or non-blocking
  - **Blocking** is considered **synchronous**
  - **Blocking send** has the sender block until the message is received
  - **Blocking receive** has the receiver block until a message is available
- 
- -

**Non-blocking** is considered **asynchronous**

**Non-blocking** send has the sender send the message and continue

**Non-blocking** receive has the receiver receive a valid message or null

### **Buffering**

Queue of messages attached to the link; implemented in one of three ways

1. Zero capacity – 0 messages

Sender must wait for receiver (rendezvous)

2. Bounded capacity – finite length of  $n$  messages

Sender must wait if link full

3. Unbounded capacity – infinite length

Sender never waits

### **Examples of IPC Systems - POSIX**

- POSIX Shared Memory
- Process first creates shared memory segment
- `segment id = shmget(IPC PRIVATE, size, S_IRUSR | S_IWUSR);`
- Process wanting access to that shared memory must attach to it
- `shared memory = (char *) shmat(id, NULL, 0);`
- Now the process could write to the shared memory
- `printf(shared memory, "Writing to shared memory");`
- When done a process can detach the shared memory from its address space
- `shmdt(shared memory);`

### **Examples of IPC Systems - Mach**

- Mach communication is message based
- Even system calls are messages
- Each task gets two mailboxes at creation- Kernel and Notify
- Only three system calls needed for message transfer
- `msg_send()`, `msg_receive()`, `msg_rpc()`
- Mailboxes needed for communication, created via
- `port_allocate()`

### **Examples of IPC Systems – Windows XP**

- Message-passing centric via local procedure call (LPC) facility
- Only works between processes on the same system
- Uses ports (like mailboxes) to establish and maintain communication channels
- Communication works as follows:

The client opens a handle to the subsystem's connection port object

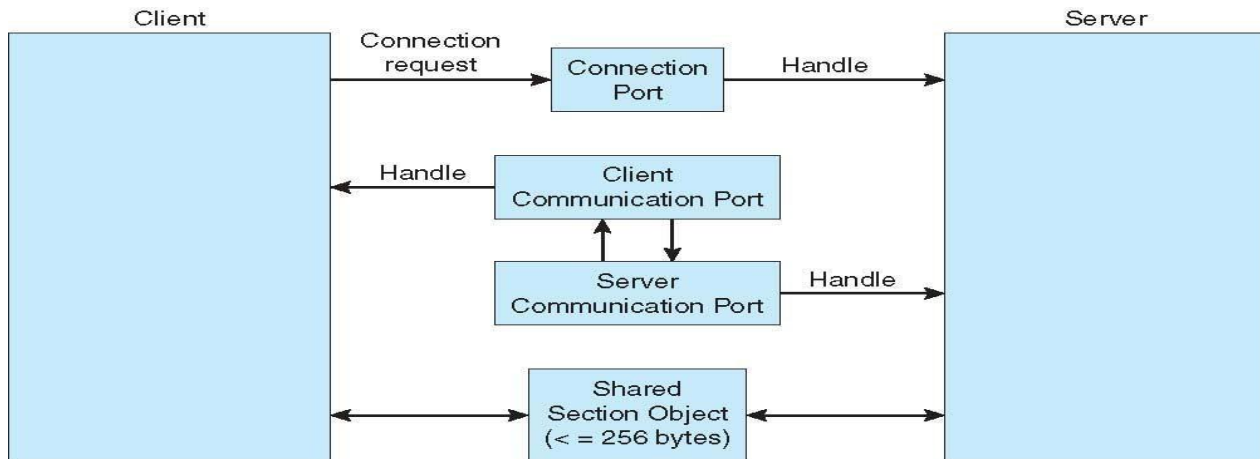
The client sends a connection request

The server creates two private communication ports and returns the handle to one of them to the client

The client and server use the corresponding port handle to send messages or callbacks and to listen for replies

---

## Local Procedure Calls in Windows XP



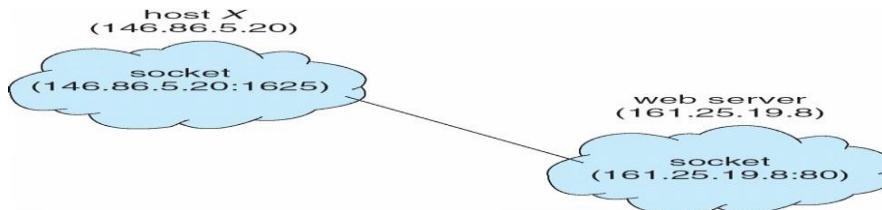
## Communications in Client-Server Systems

- Sockets
- Remote Procedure Calls
- Remote Method Invocation (Java)

### Sockets

- A socket is defined as an *endpoint for communication*
- Concatenation of IP address and port
- The socket **161.25.19.8:1625** refers to port **1625** on host **161.25.19.8**
- Communication consists between a pair of sockets

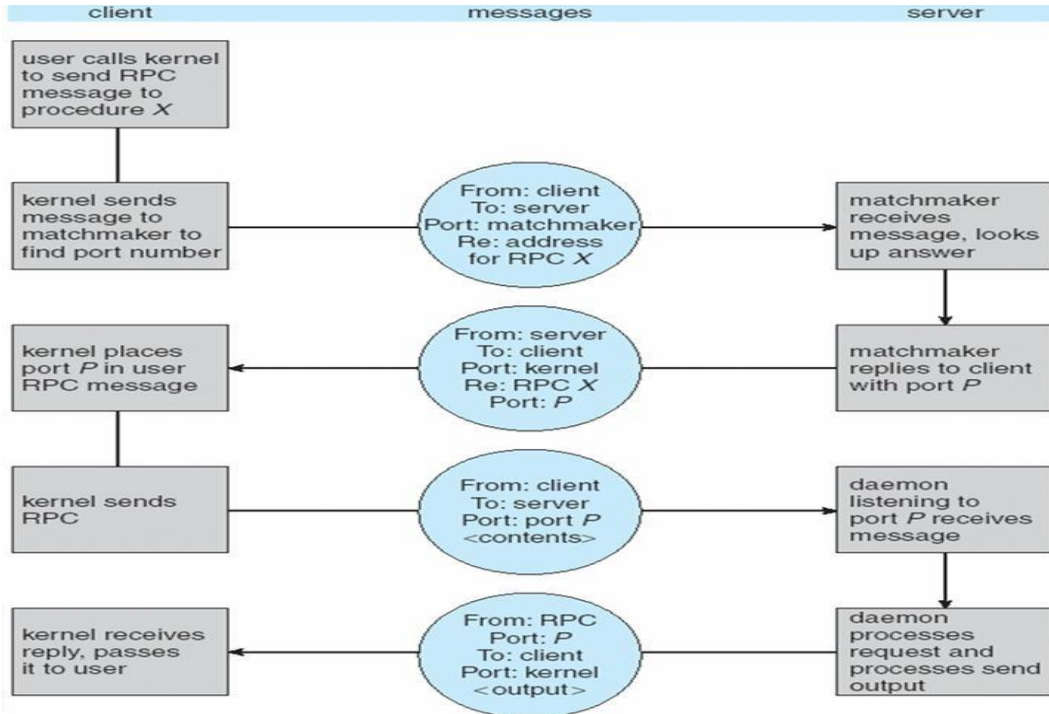
### Socket Communication



## Remote Procedure Calls

- Remote procedure call (RPC) abstracts procedure calls between processes on networked systems
- **Stubs** – client-side proxy for the actual procedure on the server
- The client-side stub locates the server and *marshalls* the parameters
- The server-side stub receives this message, unpacks the marshalled parameters, and performs the procedure on the server

## Execution of RPC

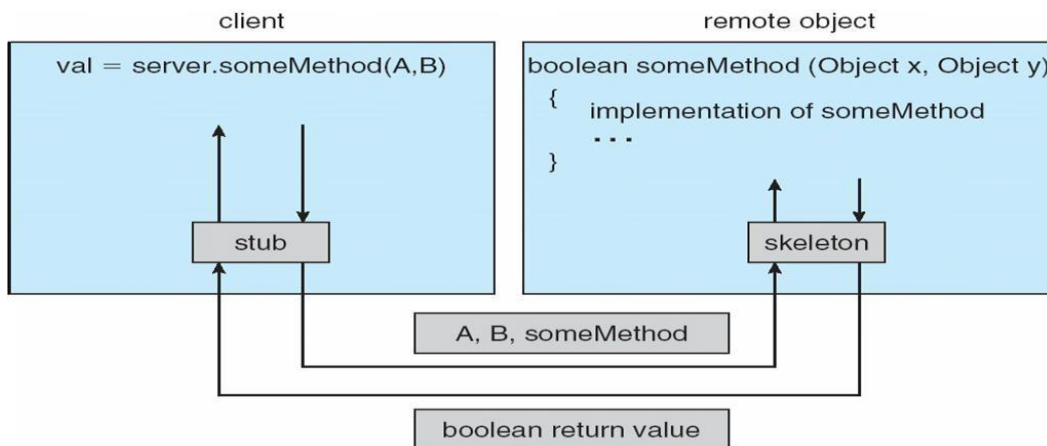


## Remote Method Invocation

- Remote Method Invocation (RMI) is a Java mechanism similar to RPCs
- RMI allows a Java program on one machine to invoke a method on a remote object



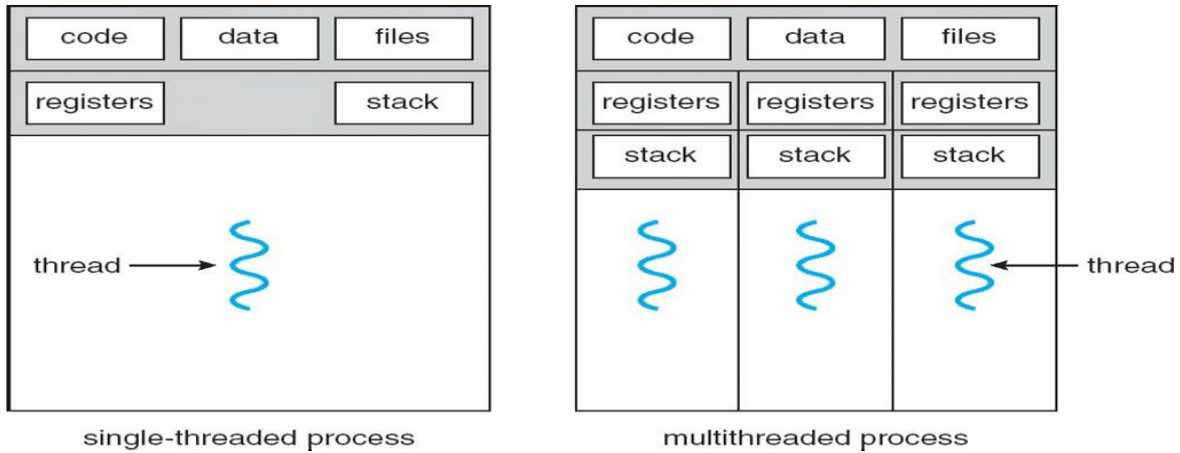
## Marshalling Parameter



## Threads

- To introduce the notion of a thread — a fundamental unit of CPU utilization that forms the basis of multithreaded computer systems
- To discuss the APIs for the Pthreads, Win32, and Java thread libraries
- To examine issues related to multithreaded programming

## Single and Multithreaded Processes



## Benefits

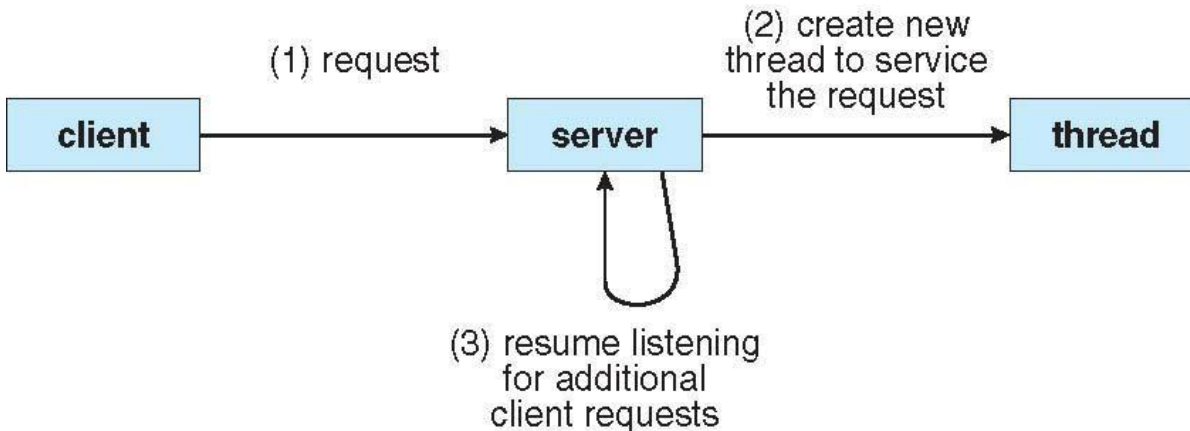
- Responsiveness
- Resource Sharing
- Economy
- Scalability

## Multicore Programming

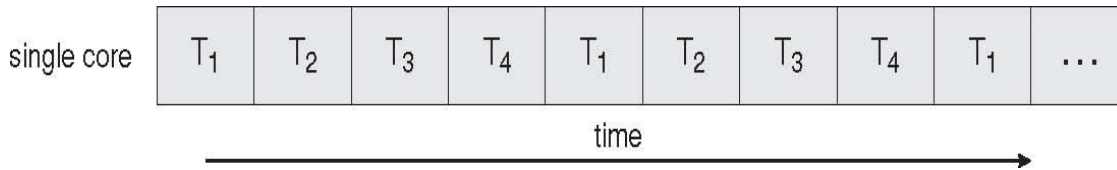
Multicore systems putting pressure on programmers, challenges include

- **Dividing activities**
- **Balance**
- **Data splitting**
- **Data dependency**
- **Testing and debugging**

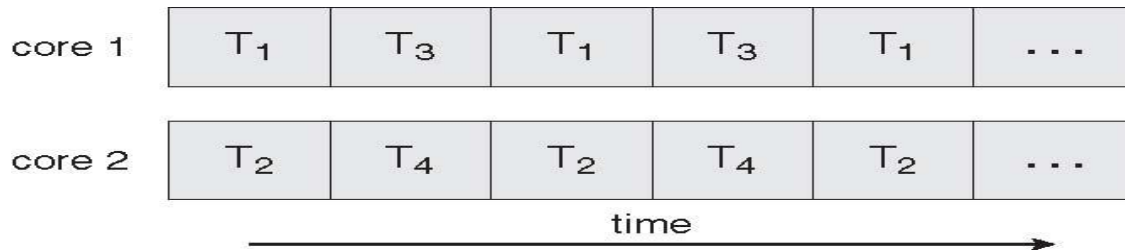
## Multithreaded Server Architecture



## Concurrent Execution on a Single-core System



## Parallel Execution on a Multicore System



## User Threads

- Thread management done by user-level threads library
- Three primary thread libraries:
  - POSIX Pthreads
  - Win32 threads
  - Java threads

## Kernel Threads

Supported by the Kernel

Examples

- Windows XP/2000
- Solaris
- Linux
- Tru64 UNIX
- Mac OS X

## Multithreading Models

- Many-to-One
- One-to-One
- Many-to-Many

## Many-to-One

Many user-level threads mapped to single kernel thread

Examples:

- Solaris Green Threads
- GNU Portable Threads

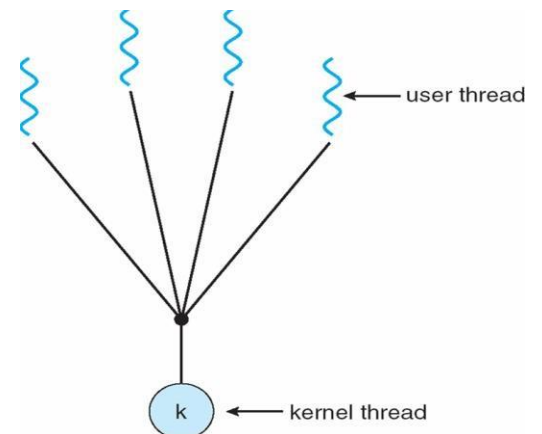
## One-to-One

Each user-level thread maps to kernel thread

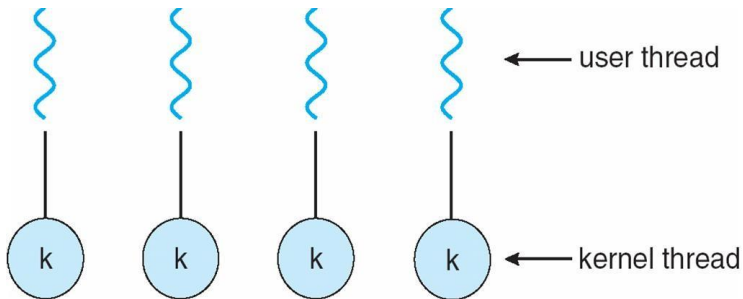
Examples

Windows NT/XP/2000

Linux



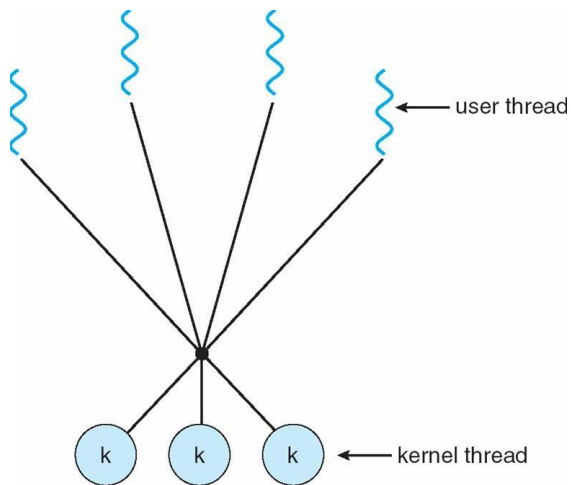
Solaris 9 and later



### Many-to-Many Model

- Allows many user level threads to be mapped to many kernel threads
- Allows the operating system to create a sufficient number of kernel threads
- Solaris prior to version 9

Windows NT/2000 with the *ThreadFiber* package

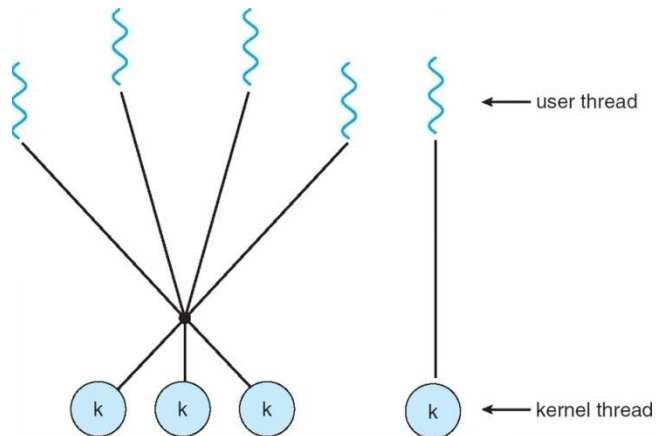


### Two-level Model

Similar to M:M, except that it allows a user thread to be **bound** to kernel thread

Examples

- IRIX
- HP-UX
- Tru64 UNIX
- Solaris 8 and earlier



## Thread Libraries

- Thread library provides programmer with API for creating and managing threads
- Two primary ways of implementing
- Library entirely in user space
- Kernel-level library supported by the OS

## Pthreads

- May be provided either as user-level or kernel-level
- A POSIX standard (IEEE 1003.1c) API for thread creation and synchronization
- API specifies behavior of the thread library, implementation is up to development of the library
- Common in UNIX operating systems (Solaris, Linux, Mac OS X)

## Java Threads

- Java threads are managed by the JVM
- Typically implemented using the threads model provided by underlying OS
- Java threads may be created by:
  - !Extending Thread class
  - !Implementing the Runnable interface

## Threading Issues

- Semantics of **fork()** and **exec()** system calls
- Thread cancellation of target thread
- Asynchronous or deferred
- Signal handling
- Thread pools
- Thread-specific data
- Scheduler activations

## Thread Cancellation

- Terminating a thread before it has finished
- Two general approaches:
- **Asynchronous cancellation** terminates the target thread immediately
- **Deferred cancellation** allows the target thread to periodically check if it should be cancelled

## Signal Handling

- Signals are used in UNIX systems to notify a process that a particular event has occurred
- A signal handler is used to process signals
- 1.Signal is generated by particular event
- 2.Signal is delivered to a process
- 3.Signal is handled
- Options:
- Deliver the signal to the thread to which the signal applies
- Deliver the signal to every thread in the process
- Deliver the signal to certain threads in the process
- Assign a specific thread to receive all signals for the process

## Thread Pools

- Create a number of threads in a pool where they await work
  - Advantages:
  - Usually slightly faster to service a request with an existing thread than create a new thread
  - Allows the number of threads in the application(s) to be bound to the size of the pool
-



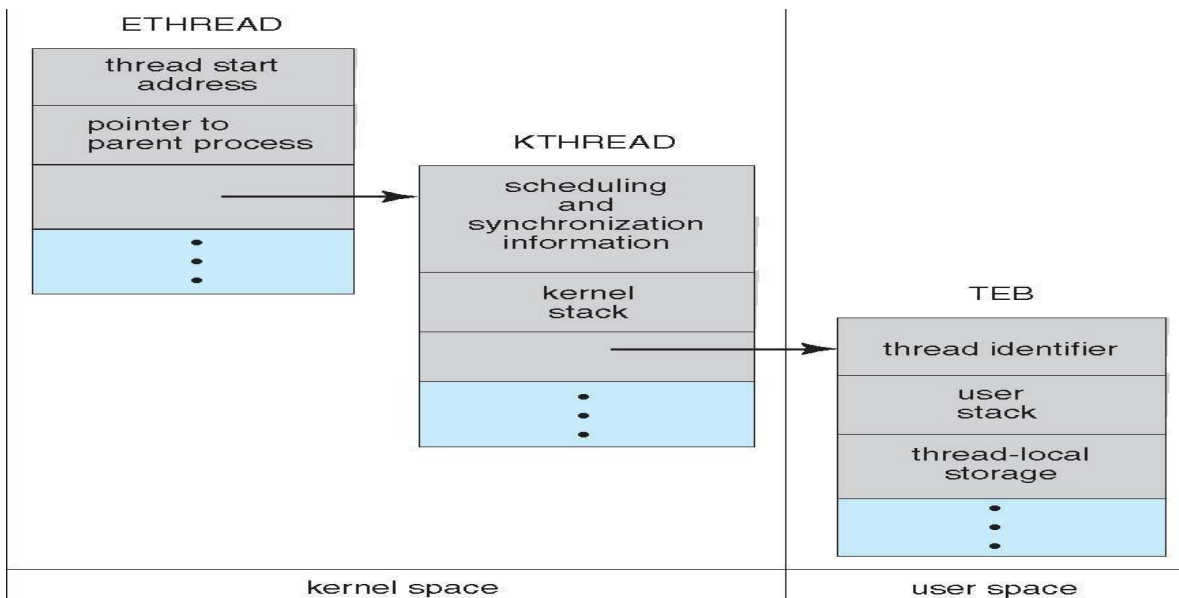
## Thread Specific Data

- Allows each thread to have its own copy of data
- Useful when you do not have control over the thread creation process (i.e., when using a thread pool)

## Scheduler Activations

- Both M:M and Two-level models require communication to maintain the appropriate number of kernel threads allocated to the application
- Scheduler activations provide upcalls - a communication mechanism from the kernel to the thread library
- This communication allows an application to maintain the correct number kernel threads

## Windows XP Threads



Implements the one-to-one mapping, kernel-level

- Each thread contains
- A thread id
- Register set
- Separate user and kernel stacks
- Private data storage area
- The register set, stacks, and private storage area are known as the context of the threads
- The primary data structures of a thread include:
- ETHREAD (executive thread block)
- KTHREAD (kernel thread block)
- TEB (thread environment block)

## Linux Threads

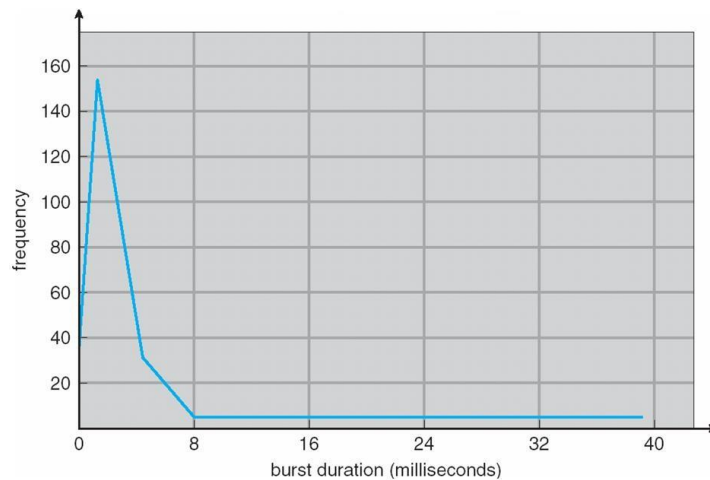
| flag          | meaning                            |
|---------------|------------------------------------|
| CLONE_FS      | File-system information is shared. |
| CLONE_VM      | The same memory space is shared.   |
| CLONE_SIGHAND | Signal handlers are shared.        |
| CLONE_FILES   | The set of open files is shared.   |

- Linux refers to them as *tasks* rather than *threads*
- Thread creation is done through **clone()** system call
- **clone()** allows a child task to share the address space of the parent task (process)

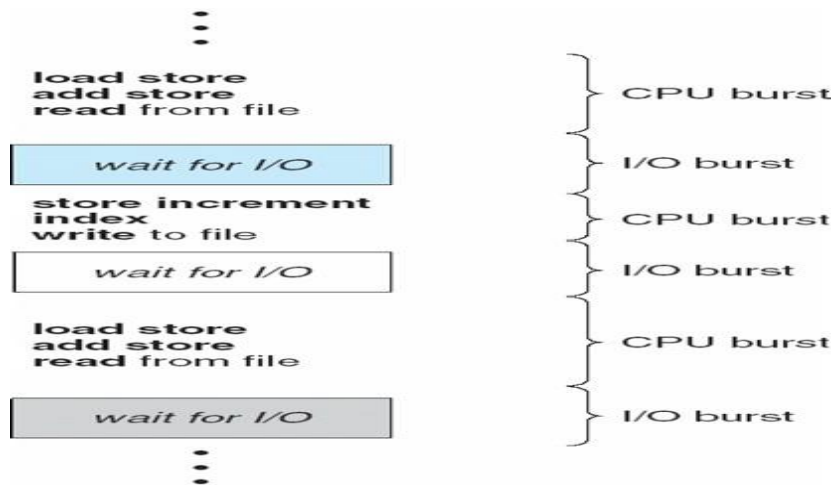
### CPU Scheduling

- To introduce CPU scheduling, which is the basis for multiprogrammed operating systems
- To describe various CPU-scheduling algorithms
- To discuss evaluation criteria for selecting a CPU-scheduling algorithm for a particular system
- Maximum CPU utilization obtained with multiprogramming
- CPU-I/O Burst Cycle – Process execution consists of a *cycle* of CPU execution and I/O wait
- **CPU burst** distribution

### Histogram of CPU-burst Times



## Alternating Sequence of CPU And I/O Bursts



### CPU Scheduler

Selects from among the processes in memory that are ready to execute, and allocates the CPU to one of them  
CPU scheduling decisions may take place when a process:

1. Switches from running to waiting state
2. Switches from running to ready state
3. Switches from waiting to ready
4. Terminates

Scheduling under 1 and 4 is **nonpreemptive**

All other scheduling is **preemptive**

### Dispatcher

- Dispatcher module gives control of the CPU to the process selected by the short-term scheduler; this involves:
  - switching context
  - switching to user mode
  - jumping to the proper location in the user program to restart that program
  - **Dispatch latency** – time it takes for the dispatcher to stop one process and start another running

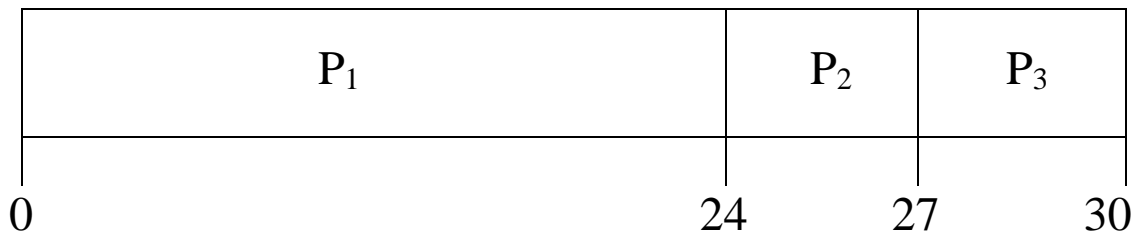
### Scheduling Criteria

- **CPU utilization** – keep the CPU as busy as possible
  - **Throughput** – # of processes that complete their execution per time unit
  - **Turnaround time** – amount of time to execute a particular process
  - **Waiting time** – amount of time a process has been waiting in the ready queue
  - **Response time** – amount of time it takes from when a request was submitted until the first response is produced, not output (for time-sharing environment)
  - Max CPU utilization
  - Max throughput
  - Min turnaround time
  - Min waiting time
  - Min response time
-

### First-Come, First-Served (FCFS) Scheduling

| Process | Burst Time |
|---------|------------|
| P1      | 24         |
| P2      | 3          |
| P3      | 3          |

Suppose that the processes arrive in the order: P1, P2, P3  
 The Gantt Chart for the schedule is:

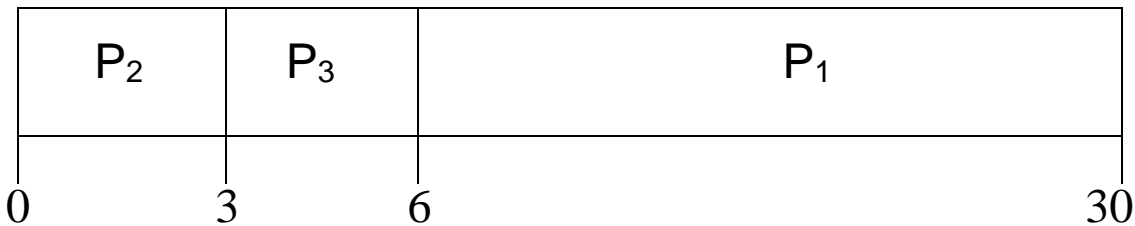


Waiting time for P1 = 0; P2 = 24; P3 = 27  
 Average waiting time:  $(0 + 24 + 27)/3 = 17$   
 Suppose that the processes arrive in the order

P2, P3, P1

The Gantt chart for the schedule is: nnnnWaiting time for P1 = 6; P2 = 0; P3 = 3  
 Average waiting time:  $(6 + 0 + 3)/3 = 3$

Much better than previous case  
 Convoy effect short process behind long process



### Shortest-Job-First (SJF) Scheduling

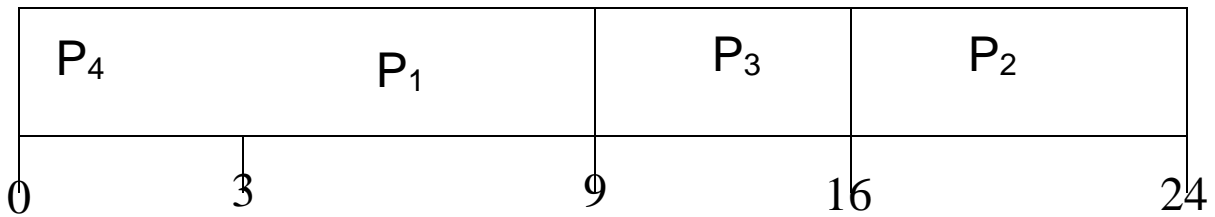
- Associate with each process the length of its next CPU burst. Use these lengths to schedule the process with the shortest time
- SJF is optimal – gives minimum average waiting time for a given set of processes

The difficulty is knowing

| Process | Arrival Time | Burst Time |
|---------|--------------|------------|
| P1      | 0.0          | 6          |
| P2      | 2.0          | 8          |
| P3      | 4.0          | 7          |
| P4      | 5.0          | 3          |

## SJF scheduling chart

average waiting time =  $(3 + 16 + 9 + 0) / 4 = 7$  the length of the next CPU request

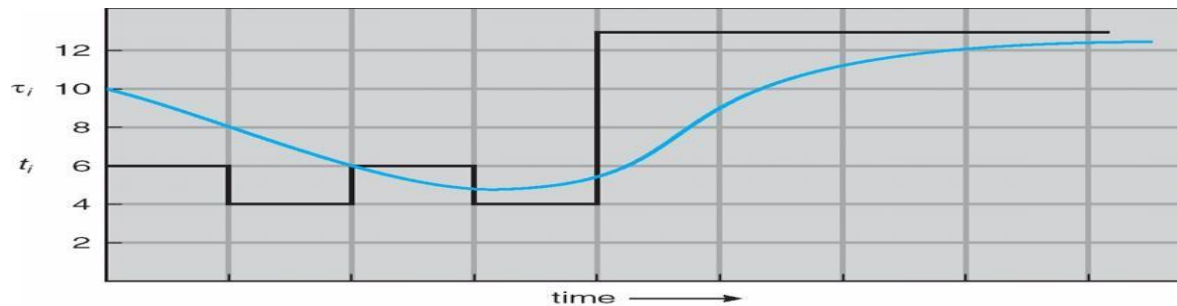


### Determining Length of Next CPU Burst

1.  $t_n$  = actual length of  $n^{th}$  CPU burst
2.  $\tau_{n+1}$  = predicted value for the next CPU burst
3.  $\alpha, 0 \leq \alpha \leq 1$
4. Define :

- Can only estimate the length
- Can be done by using the length of previous CPU bursts, using exponential averaging

### Prediction of the Length of the Next CPU Burst



|                      |    |   |   |   |    |    |    |     |
|----------------------|----|---|---|---|----|----|----|-----|
| CPU burst ( $t_i$ )  | 6  | 4 | 6 | 4 | 13 | 13 | 13 | ... |
| "guess" ( $\tau_i$ ) | 10 | 8 | 6 | 6 | 9  | 11 | 12 | ... |

### Examples of Exponential Averaging

$\alpha = 0$

$$t_{n+1} = t_n$$

Recent history does not count

$\alpha = 1$

$$t_{n+1} = \alpha t_n$$

Only the actual last CPU burst counts

If we expand the formula, we get:

$$t_{n+1} = \alpha t_n + (1 - \alpha) t_{n-1} + \dots$$

$$+ (1 - \alpha)^j t_{n-j} + \dots$$

$$+(1 - a)^{n+1} t_0$$

Since both  $a$  and  $(1 - a)$  are less than or equal to 1, each successive term has less weight than its predecessor

### Priority Scheduling

- A priority number (integer) is associated with each process
- The CPU is allocated to the process with the highest priority (smallest integer ° highest priority)
- Preemptive
- nonpreemptive
- SJF is a priority scheduling where priority is the predicted next CPU burst time
- Problem ° **Starvation** – low priority processes may never execute
- Solution ° **Aging** – as time progresses increase the priority of the process

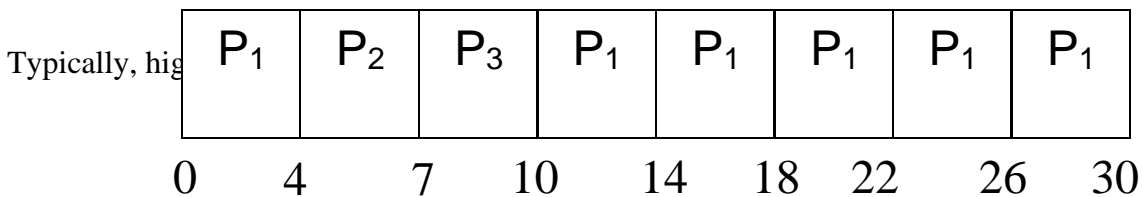
### Round Robin (RR)

- Each process gets a small unit of CPU time (*time quantum*), usually 10-100 milliseconds. After this time has elapsed, the process is preempted and added to the end of the ready queue.
- If there are  $n$  processes in the ready queue and the time quantum is  $q$ , then each process gets  $1/n$  of the CPU time in chunks of at most  $q$  time units at once. No process waits more than  $(n-1)q$  time units.
- Performance
- $q$  large P FIFO
- $q$  small P  $q$  must be large with respect to context switch, otherwise overhead is too high

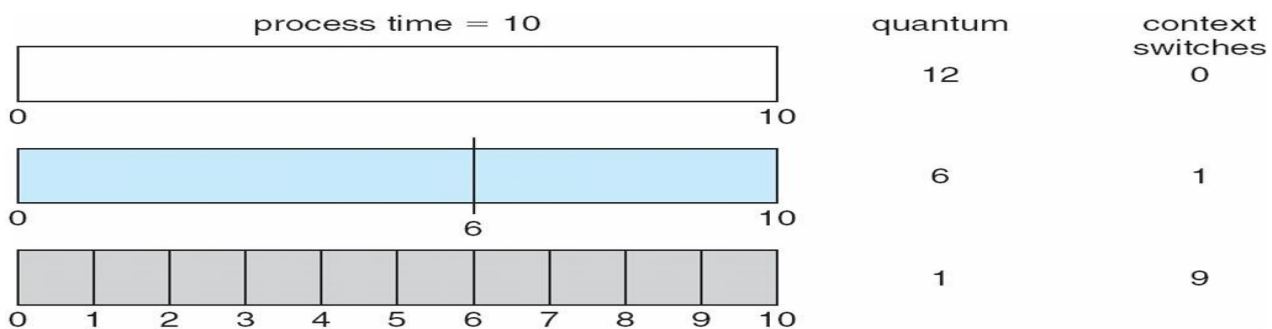
### Example of RR with Time Quantum = 4

| Process | Burst Time |
|---------|------------|
| $P_1$   | 24         |
| $P_2$   | 3          |
| $P_3$   | 3          |

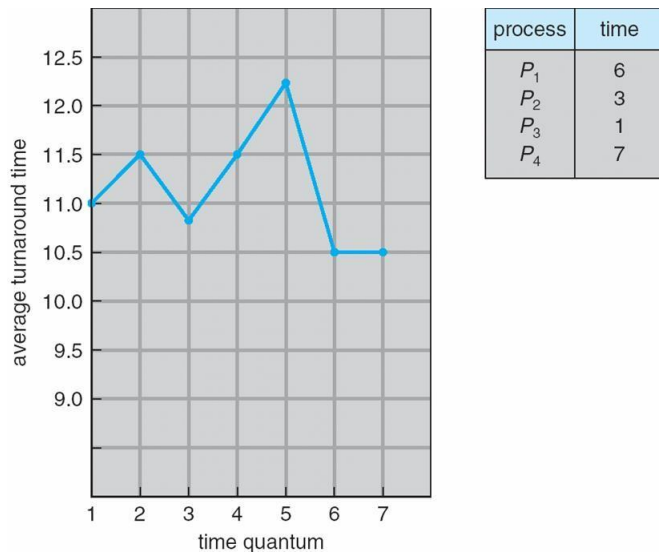
The Gantt chart is:



### Time Quantum and Context Switch Time



## Turnaround Time Varies With The Time Quantum

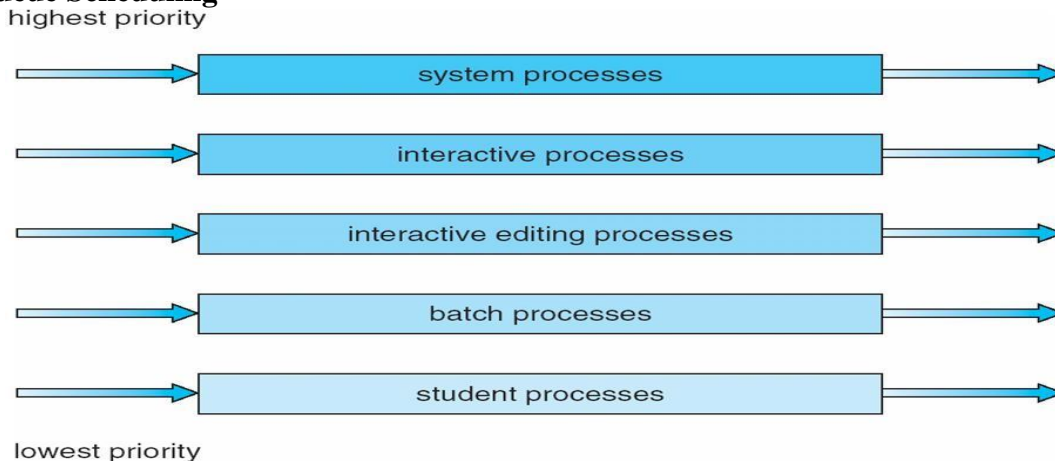


### Multilevel Queue

- Ready queue is partitioned into separate queues:  
foreground (interactive)  
background (batch)
- Each queue has its own scheduling algorithm
- foreground – RR
- background – FCFS
- Scheduling must be done between the queues
- Fixed priority scheduling; (i.e., serve all from foreground then from background). Possibility of starvation.
- Time slice – each queue gets a certain amount of CPU time which it can schedule amongst its processes;  
i.e., 80% to foreground in RR

20% to background in FCFS

### Multilevel Queue Scheduling



## Multilevel Feedback Queue

- A process can move between the various queues; aging can be implemented this way
- Multilevel-feedback-queue scheduler defined by the following parameters:
- number of queues
- scheduling algorithms for each queue
- method used to determine when to upgrade a process
- method used to determine when to demote a process

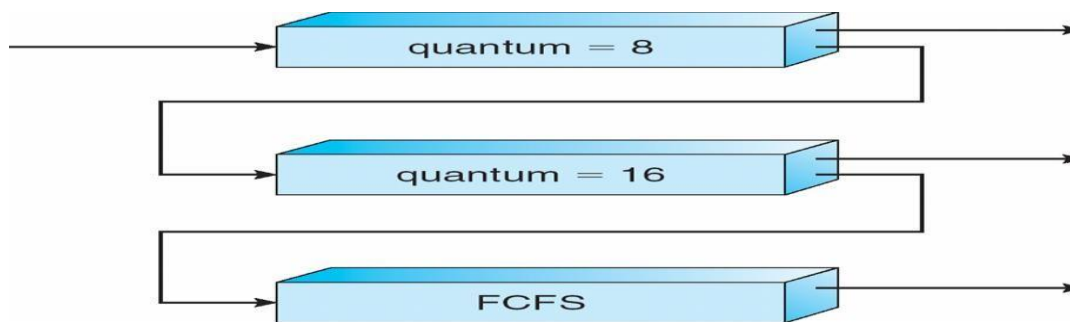
method used to determine which queue a process will enter when that process needs service

## Example of Multilevel Feedback Queue

Three queues:

- $Q_0$  – RR with time quantum 8 milliseconds
- $Q_1$  – RR time quantum 16 milliseconds
- $Q_2$  – FCFS
- Scheduling
- A new job enters queue  $Q_0$  which is served FCFS. When it gains CPU, job receives 8 milliseconds. If it does not finish in 8 milliseconds, job is moved to queue  $Q_1$ .
- At  $Q_1$  job is again served FCFS and receives 16 additional milliseconds. If it still does not complete, it is preempted and moved to queue  $Q_2$ .

## Multilevel Feedback Queues



## Thread Scheduling

- Distinction between user-level and kernel-level threads
- Many-to-one and many-to-many models, thread library schedules user-level threads to run on LWP
- Known as **process-contention scope (PCS)** since scheduling competition is within the process
- Kernel thread scheduled onto available CPU is **system-contention scope (SCS)** – competition among all threads in system

## Pthread Scheduling

- API allows specifying either PCS or SCS during thread creation
- PTHREAD SCOPE PROCESS schedules threads using PCS scheduling
- PTHREAD SCOPE SYSTEM schedules threads using SCS scheduling.

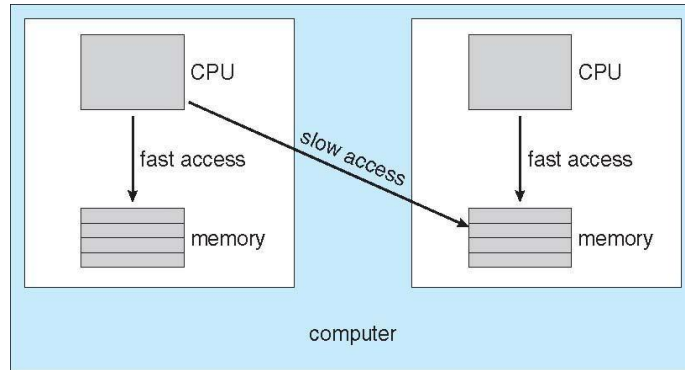
## Multiple-Processor Scheduling

- CPU scheduling more complex when multiple CPUs are available
  - **Homogeneous processors** within a multiprocessor
  - **Asymmetric multiprocessing** – only one processor accesses the system data structures, alleviating the need for data sharing
-



- **Symmetric multiprocessing (SMP)** – each processor is self-scheduling, all processes in common ready queue, or each has its own private queue of ready processes
- **Processor affinity** – process has affinity for processor on which it is currently running
- **soft affinity**
- **hard affinity**

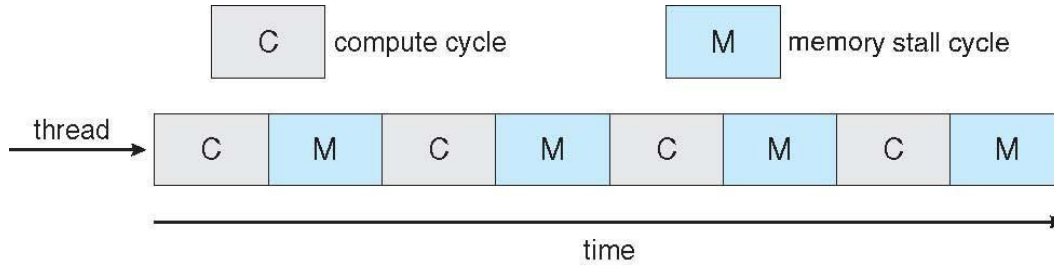
### NUMA and CPU Scheduling



## Multicore Processors

- Recent trend to place multiple processor cores on same physical chip
- Faster and consume less power
- Multiple threads per core also growing
- Takes advantage of memory stall to make progress on another thread while memory retrieve happens

## Multithreaded Multicore System



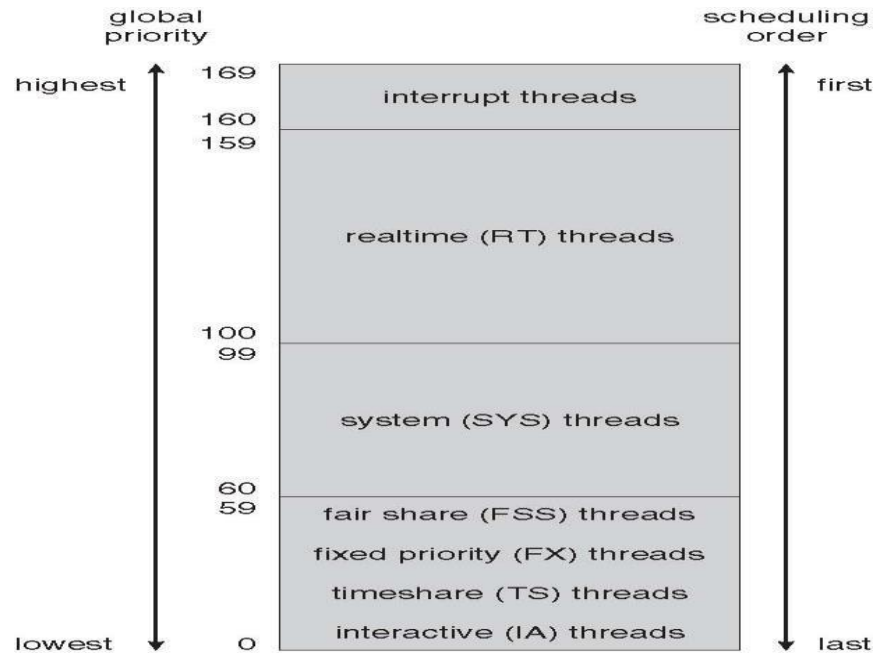
## Operating System Examples

- Solaris scheduling
- Windows XP scheduling
- Linux scheduling

## Solaris Dispatch Table

| priority | time quantum | time quantum expired | return from sleep |
|----------|--------------|----------------------|-------------------|
| 0        | 200          | 0                    | 50                |
| 5        | 200          | 0                    | 50                |
| 10       | 160          | 0                    | 51                |
| 15       | 160          | 5                    | 51                |
| 20       | 120          | 10                   | 52                |
| 25       | 120          | 15                   | 52                |
| 30       | 80           | 20                   | 53                |
| 35       | 80           | 25                   | 54                |
| 40       | 40           | 30                   | 55                |
| 45       | 40           | 35                   | 56                |
| 50       | 40           | 40                   | 58                |
| 55       | 40           | 45                   | 58                |
| 59       | 20           | 49                   | 59                |

## Solaris Scheduling



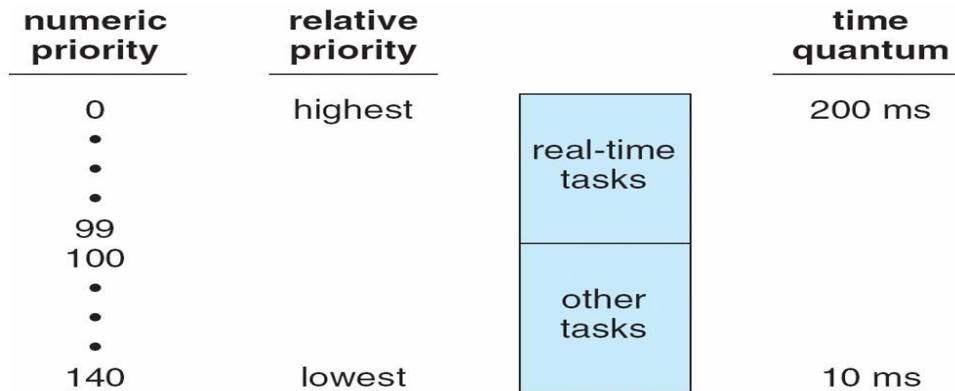
## Windows XP Priorities

|               | real-time | high | above normal | normal | below normal | idle priority |
|---------------|-----------|------|--------------|--------|--------------|---------------|
| time-critical | 31        | 15   | 15           | 15     | 15           | 15            |
| highest       | 26        | 15   | 12           | 10     | 8            | 6             |
| above normal  | 25        | 14   | 11           | 9      | 7            | 5             |
| normal        | 24        | 13   | 10           | 8      | 6            | 4             |
| below normal  | 23        | 12   | 9            | 7      | 5            | 3             |
| lowest        | 22        | 11   | 8            | 6      | 4            | 2             |
| idle          | 16        | 1    | 1            | 1      | 1            | 1             |

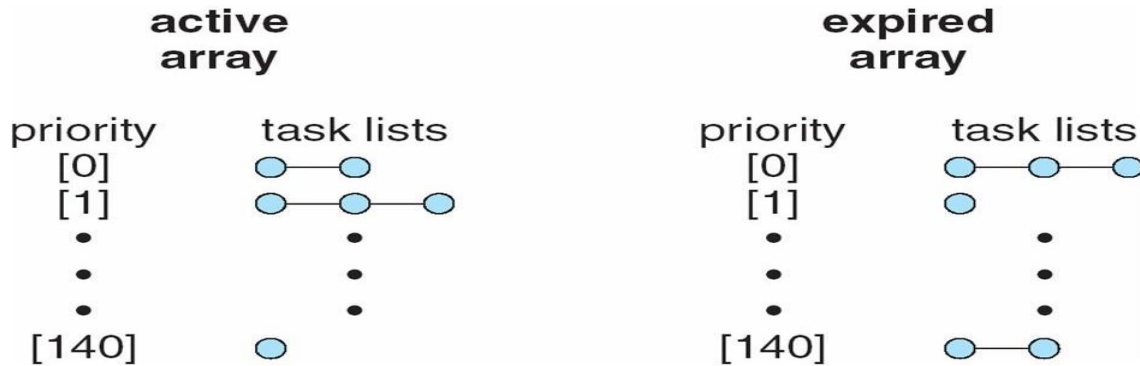
## Linux Scheduling

- Constant order  $O(1)$  scheduling time
- Two priority ranges: time-sharing and real-time
- **Real-time** range from 0 to 99 and **nice** value from 100 to 140

## Priorities and Time-slice length



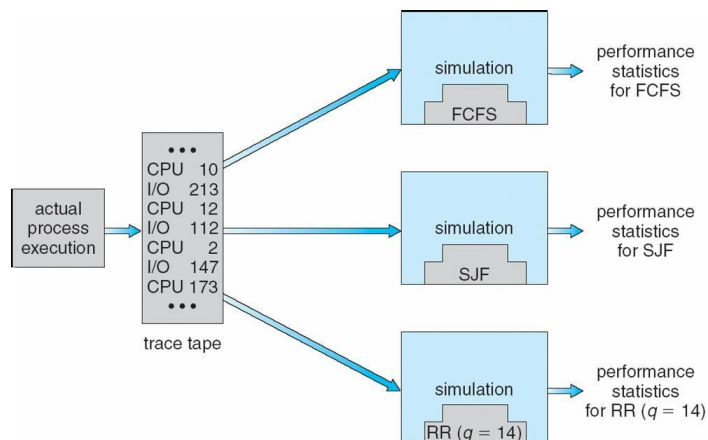
## List of Tasks Indexed According to Priorities



## Algorithm Evaluation

- Deterministic modeling – takes a particular predetermined workload and defines the performance of each algorithm for that workload
- Queuing models
- Implementation

## Evaluation of CPU schedulers by Simulation



# UNIT -3



# CONCURRENCY

## Process Synchronization

- To introduce the critical-section problem, whose solutions can be used to ensure the consistency of shared data
- To present both software and hardware solutions of the critical-section problem
- To introduce the concept of an atomic transaction and describe mechanisms to ensure atomicity
- Concurrent access to shared data may result in data inconsistency
- Maintaining data consistency requires mechanisms to ensure the orderly execution of cooperating processes
- Suppose that we wanted to provide a solution to the consumer-producer problem that fills all the buffers. We can do so by having an integer count that keeps track of the number of full buffers. Initially, count is set to 0. It is incremented by the producer after it produces a new buffer and is decremented by the consumer after it consumes a buffer

## Checkpoints

Log could become long, and recovery could take long

Checkpoints shorten log and recovery time.

Checkpoint scheme:

1. Output all log records currently in volatile storage to stable storage
2. Output all modified data from volatile to stable storage
3. Output a log record <checkpoint> to the log on stable storage

Now recovery only includes  $T_i$ , such that  $T_i$  started executing before the most recent checkpoint, and all transactions after  $T_i$ . All other transactions already on stable storage

## Concurrent Transactions

- Must be equivalent to serial execution – serializability
- Could perform all transactions in critical section
- Inefficient, too restrictive
- Concurrency-control algorithms provide serializability

## Serializability

- Consider two data items A and B
- Consider Transactions  $T_0$  and  $T_1$
- Execute  $T_0, T_1$  atomically
- Execution sequence called schedule
- Atomically executed transaction order called serial schedule
- For N transactions, there are  $N!$  valid serial schedules

### Schedule 1: $T_0$ then $T_1$

|  | $T_0$    | $T_1$    |
|--|----------|----------|
|  | read(A)  |          |
|  | write(A) |          |
|  | read(B)  |          |
|  | write(B) |          |
|  |          | read(A)  |
|  |          | write(A) |

## Nonserial Schedule

- Nonserial schedule allows overlapped execute
- Resulting execution not necessarily incorrect
- Consider schedule S, operations  $O_i, O_j$
- Conflict if access same data item, with at least one write
- If  $O_i, O_j$  consecutive and operations of different transactions &  $O_i$  and  $O_j$  don't conflict
- Then  $S'$  with swapped order  $O_j O_i$  equivalent to S
- If S can become  $S'$  via swapping nonconflicting operations
- S is conflict serializable
- Schedule 2: Concurrent Serializable Schedule

| $T_0$    | $T_1$    |
|----------|----------|
| read(A)  |          |
| write(A) |          |
|          | read(A)  |
|          | write(A) |
| read(B)  |          |
| write(B) |          |
|          | read(B)  |
|          | write(B) |

## Locking Protocol

- Ensure serializability by associating lock with each data item
- Follow locking protocol for access control
- Locks
- Shared –  $T_i$  has shared-mode lock (S) on item Q,  $T_i$  can read Q but not write Q
- Exclusive –  $T_i$  has exclusive-mode lock (X) on Q,  $T_i$  can read and write Q
- Require every transaction on item Q acquire appropriate lock
- If lock already held, new request may have to wait
- Similar to readers-writers algorithm

## Two-phase Locking Protocol

- Generally ensures conflict serializability
- Each transaction issues lock and unlock requests in two phases
- Growing – obtaining locks
- Shrinking – releasing locks
- Does not prevent deadlock

## Timestamp-based Protocols

- Select order among transactions in advance – timestamp-ordering
- Transaction  $T_i$  associated with timestamp  $TS(T_i)$  before  $T_i$  starts
- $TS(T_i) < TS(T_j)$  if  $T_i$  entered system before  $T_j$
- TS can be generated from system clock or as logical counter incremented at each entry of transaction

- 
- -

Timestamps determine serializability order

If  $TS(T_i) < TS(T_j)$ , system must ensure produced schedule equivalent to serial schedule where  $T_i$  appears before  $T_j$

### Timestamp-based Protocol Implementation

- Data item  $Q$  gets two timestamps
- $W$ -timestamp( $Q$ ) – largest timestamp of any transaction that executed  $write(Q)$  successfully
- $R$ -timestamp( $Q$ ) – largest timestamp of successful  $read(Q)$
- Updated whenever  $read(Q)$  or  $write(Q)$  executed
- Timestamp-ordering protocol assures any conflicting read and write executed in timestamp order

Suppose  $T_i$  executes  $read(Q)$

If  $TS(T_i) < W$ -timestamp( $Q$ ),  $T_i$  needs to read value of  $Q$  that was already overwritten read operation rejected and  $T_i$  rolled back

If  $TS(T_i) \geq W$ -timestamp( $Q$ ) read executed,  $R$ -timestamp( $Q$ ) set to  $\max(R$ -timestamp( $Q$ ),  $TS(T_i)$ )

### Timestamp-ordering Protocol

Suppose  $T_i$  executes  $write(Q)$

If  $TS(T_i) < R$ -timestamp( $Q$ ), value  $Q$  produced by  $T_i$  was needed previously and  $T_i$  assumed it would never be produced Write operation rejected,  $T_i$  rolled back If  $TS(T_i) < W$ -timestamp( $Q$ ),  $T_i$  attempting to write obsolete value of  $Q$  Write operation rejected and  $T_i$  rolled back Otherwise, write executed Any rolled back transaction  $T_i$  is assigned new timestamp and restarted Algorithm ensures conflict serializability and freedom from deadlock

### Schedule Possible Under Timestamp Protocol

| $T_2$       | $T_3$        |
|-------------|--------------|
| read( $B$ ) |              |
|             | read( $B$ )  |
|             | write( $B$ ) |
| read( $A$ ) |              |
|             | read( $A$ )  |
|             | write( $A$ ) |



# UNIT -4

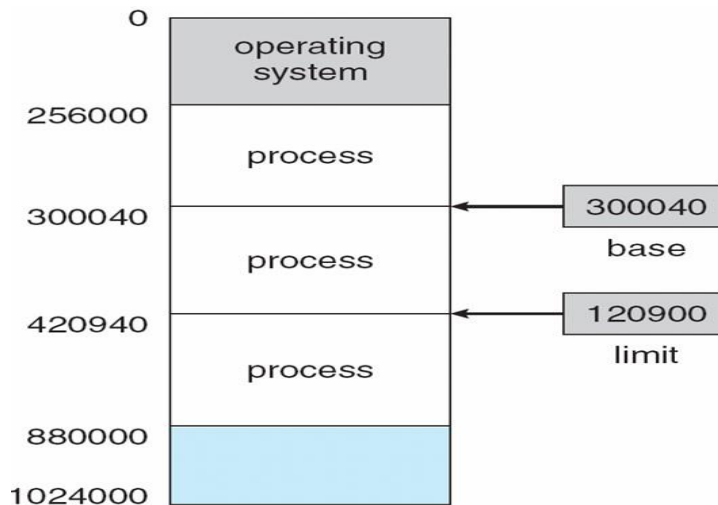


## Memory Management

- To provide a detailed description of various ways of organizing memory hardware
- To discuss various memory-management techniques, including paging and segmentation
- To provide a detailed description of the Intel Pentium, which supports both pure segmentation and segmentation with paging
- Program must be brought (from disk) into memory and placed within a process for it to be run
- Main memory and registers are only storage CPU can access directly
- Register access in one CPU clock (or less)
- Main memory can take many cycles
- **Cache** sits between main memory and CPU registers
- Protection of memory required to ensure correct operation

### Base and Limit Registers

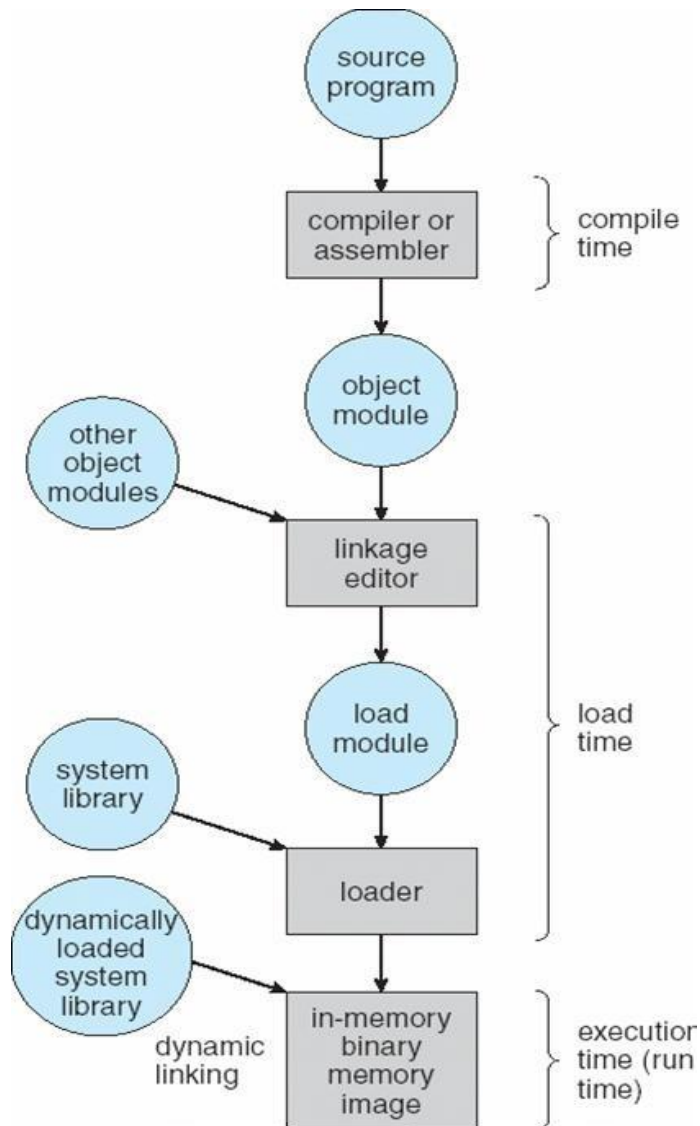
A pair of **base** and **limit** registers define the logical address space



### Binding of Instructions and Data to Memory

- Address binding of instructions and data to memory addresses can happen at three different stages
  - **Compile time:** If memory location known a priori, **absolute code** can be generated; must recompile code if starting location changes
  - **Load time:** Must generate **relocatable code** if memory location is not known at compile time
  - **Execution time:** Binding delayed until run time if the process can be moved during its execution from one memory segment to another. Need hardware support for address maps (e.g., base and limit registers)
-

## Multistep Processing of a User Program



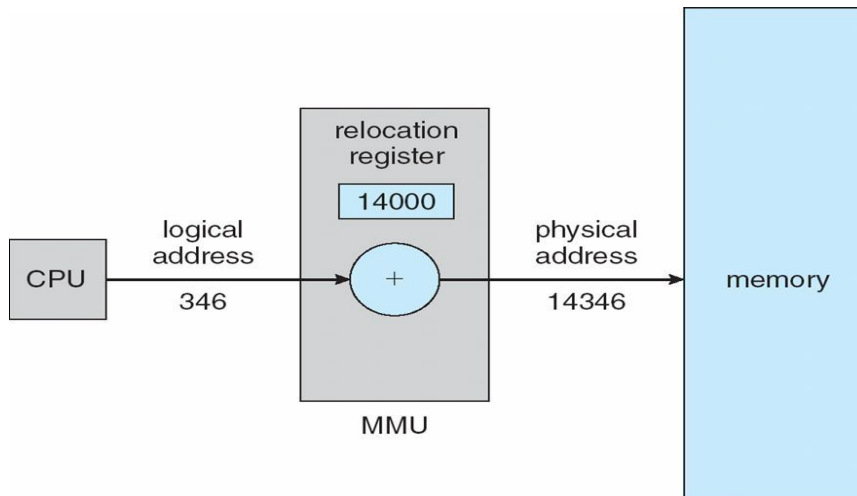
## Logical vs. Physical Address Space

- The concept of a logical address space that is bound to a separate **physical address space** is central to proper memory management
  - **Logical address** – generated by the CPU; also referred to as **virtual address**
  - **Physical address** – address seen by the memory unit
  - Logical and physical addresses are the same in compile-time and load-time address-binding schemes; logical (virtual) and physical addresses differ in execution-time address-binding scheme
-

## Memory-Management Unit (MMU)

- Hardware device that maps virtual to physical address
- In MMU scheme, the value in the relocation register is added to every address generated by a user process at the time it is sent to memory
- The user program deals with *logical* addresses; it never sees the *real* physical addresses

### Dynamic relocation using a relocation register



## Dynamic Loading

- Routine is not loaded until it is called
- Better memory-space utilization; unused routine is never loaded
- Useful when large amounts of code are needed to handle infrequently occurring cases
- No special support from the operating system is required implemented through program design

## Dynamic Linking

- Linking postponed until execution time
- Small piece of code, *stub*, used to locate the appropriate memory-resident library routine
- Stub replaces itself with the address of the routine, and executes the routine
- Operating system needed to check if routine is in processes' memory address
- Dynamic linking is particularly useful for libraries
- System also known as **shared libraries**

## Swapping

A process can be swapped temporarily out of memory to a backing store, and then brought back into memory for continued execution.

**Backing store** – fast disk large enough to accommodate copies of all memory images for all users; must provide direct access to these memory images.

**Roll out, roll in** – swapping variant used for priority-based scheduling algorithms; lower-priority process is swapped out so higher-priority process can be loaded and executed.

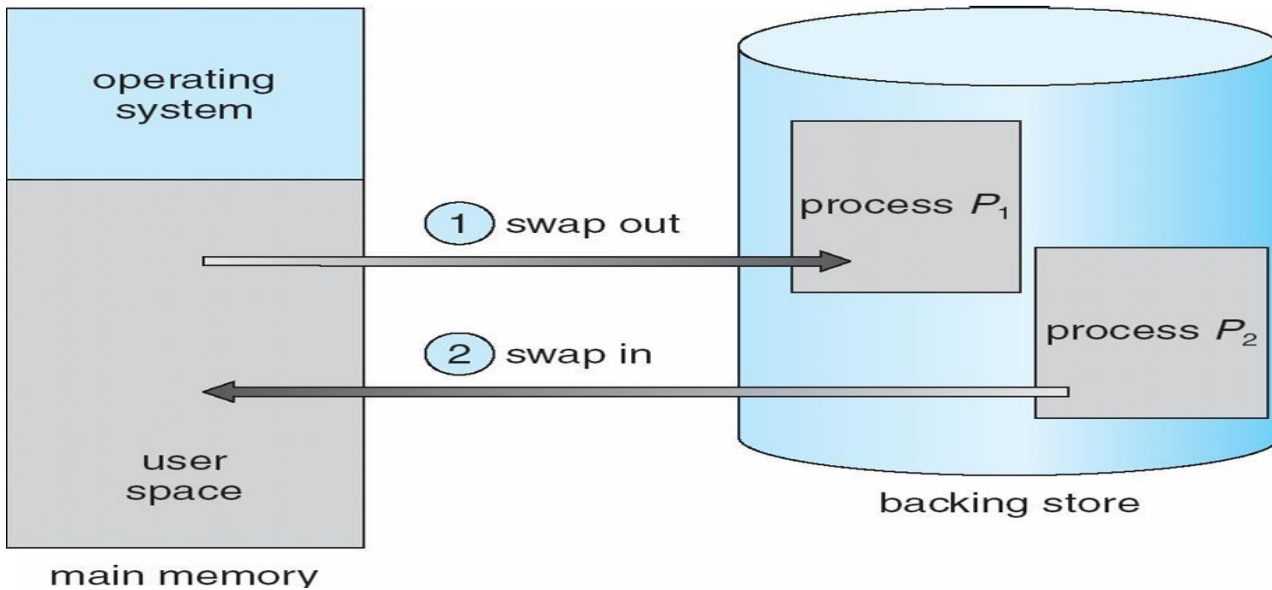
Major part of swap time is transfer time; total transfer time is directly proportional to the amount of memory swapped.

Modified versions of swapping are found on many systems (i.e., UNIX, Linux, and Windows)

System maintains a **ready queue** of ready-to-run processes which have memory images on disk

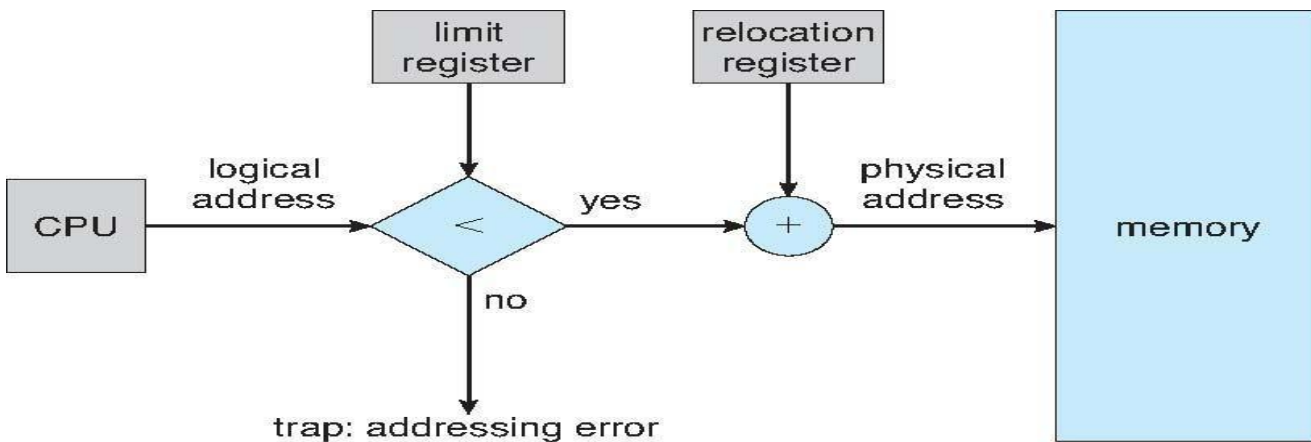
---

## Schematic View of Swapping



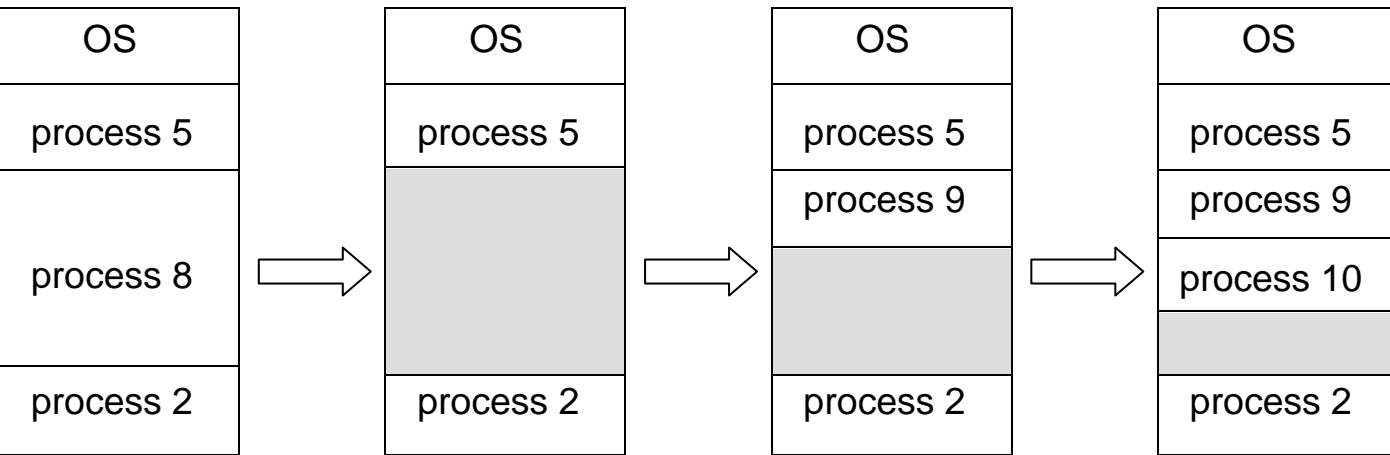
- Main memory usually into two partitions:
- Resident operating system, usually held in low memory with interrupt vector
- User processes then held in high memory
- Relocation registers used to protect user processes from each other, and from changing operating-system code and data
- Base register contains value of smallest physical address
- Limit register contains range of logical addresses – each logical address must be less than the limit register
- MMU maps logical address *dynamically*

## Hardware Support for Relocation and Limit Registers



- Multiple-partition allocation
  - Hole – block of available memory; holes of various size are scattered throughout memory
  - When a process arrives, it is allocated memory from a hole large enough to accommodate it
-

- Operating system maintains information about:
  - a) allocated partitions    b) free partitions (hole)



### Dynamic Storage-Allocation Problem

- **First-fit:** Allocate the *first* hole that is big enough
- **Best-fit:** Allocate the *smallest* hole that is big enough; must search entire list, unless ordered by size  
Produces the smallest leftover hole
- **Worst-fit:** Allocate the *largest* hole; must also search entire list  
Produces the largest leftover hole
- First-fit and best-fit better than worst-fit in terms of speed and storage utilization

### Fragmentation

- **External Fragmentation** – total memory space exists to satisfy a request, but it is not contiguous
- **Internal Fragmentation** – allocated memory may be slightly larger than requested memory; this size difference is memory internal to a partition, but not being used
- Reduce external fragmentation by **compaction**
- Shuffle memory contents to place all free memory together in one large block
- Compaction is possible *only* if relocation is dynamic, and is done at execution time.
- I/O problem
  - Latch job in memory while it is involved in I/O
  - Do I/O only into OS buffers

### Paging

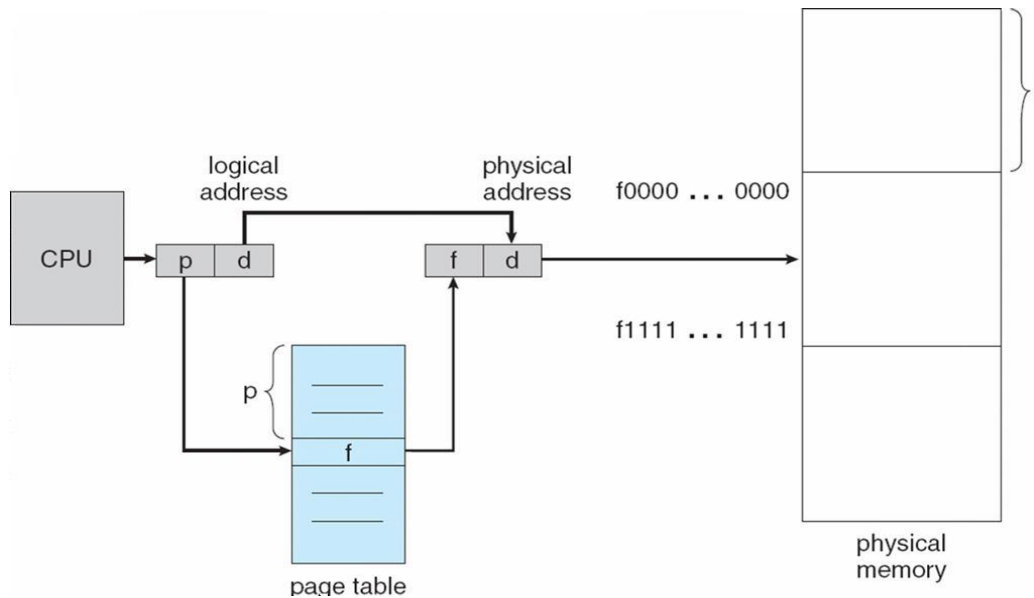
- Logical address space of a process can be noncontiguous; process is allocated physical memory whenever the latter is available
  - Divide physical memory into fixed-sized blocks called **frames** (size is power of 2, between 512 bytes and 8,192 bytes)
  - Divide logical memory into blocks of same size called **pages** Keep track of all free frames
-

- To run a program of size  $n$  pages, need to find  $n$  free frames and load program
- Set up a page table to translate logical to physical addresses
- Internal fragmentation

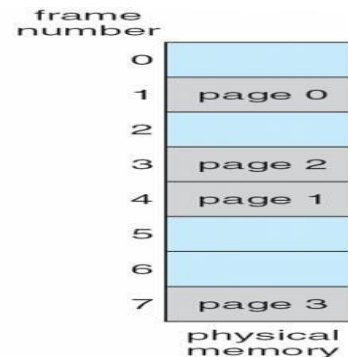
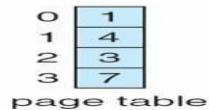
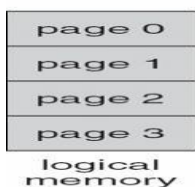
## Address Translation Scheme

- Address generated by CPU is divided into
- **Page number ( $p$ )** – used as an index into a *page table* which contains base address of each page in physical memory
- **Page offset ( $d$ )** – combined with base address to define the physical memory address that is sent to the memory unit
- For given logical address space  $2^m$  and *page size*  $2^n$

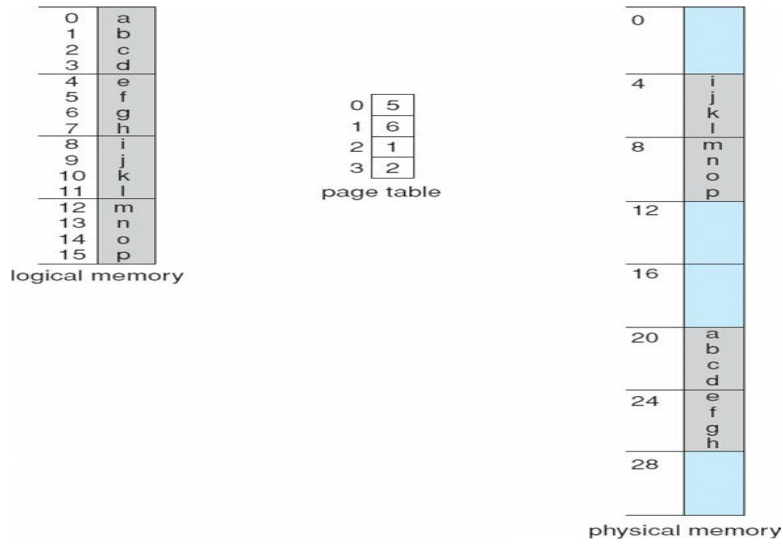
## Paging Hardware



## Paging Model of Logical and Physical Memory

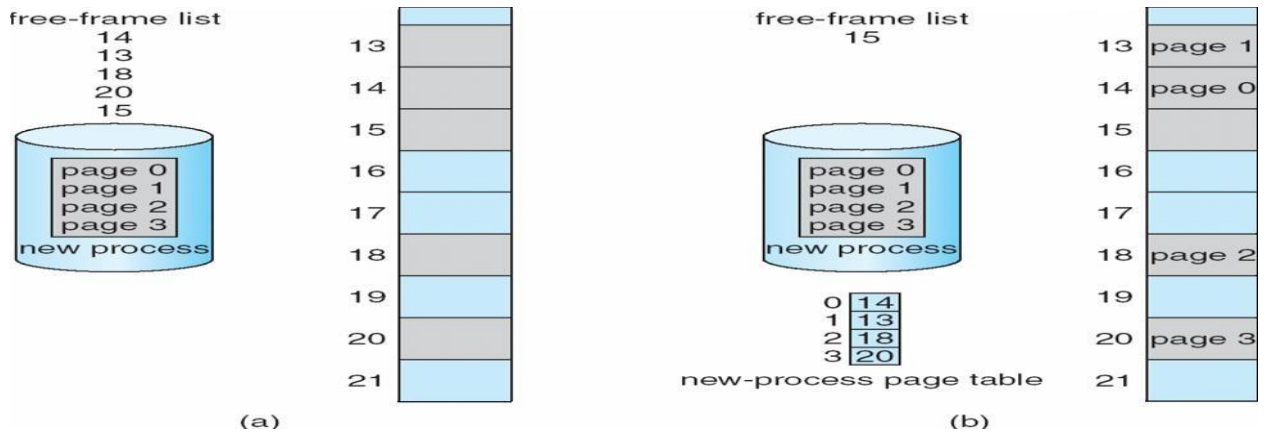


# Paging Example



## 32-byte memory and 4-byte pages

### Free Frames





## Implementation of Page Table

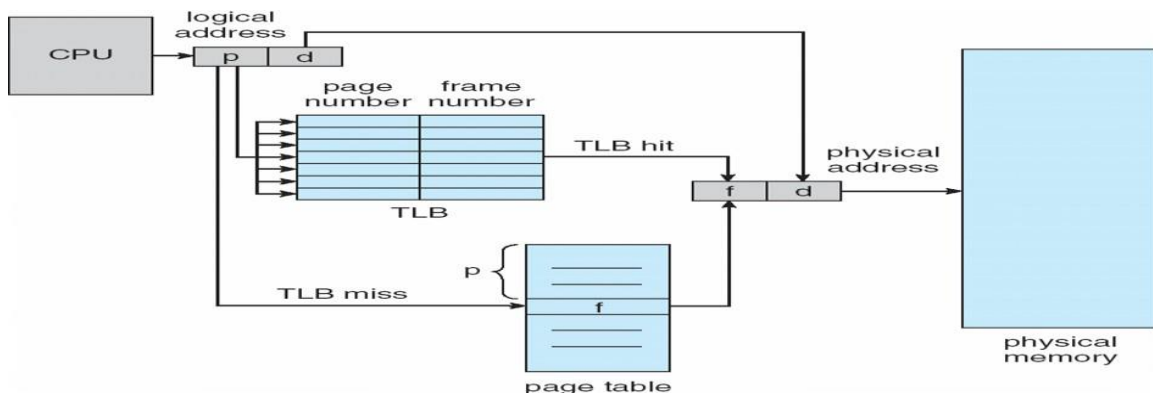
- Page table is kept in main memory
- **Page-table base register (PTBR)** points to the page table
- **Page-table length register (PRLR)** indicates size of the page table
- In this scheme every data/instruction access requires two memory accesses. One for the page table and one for the data/instruction.
- The two memory access problem can be solved by the use of a special fast-lookup hardware cache called **associative memory** or **translation look-aside buffers (TLBs)**
- Some TLBs store **address-space identifiers (ASIDs)** in each TLB entry – uniquely identifies each process to provide address-space protection for that process

## Associative Memory

- Associative memory – parallel search
- Address translation (p, d)
- If p is in associative register, get frame # out
- Otherwise get frame # from page table in memory

| Page # | Frame # |
|--------|---------|
|        |         |
|        |         |
|        |         |
|        |         |

## Paging Hardware With TLB



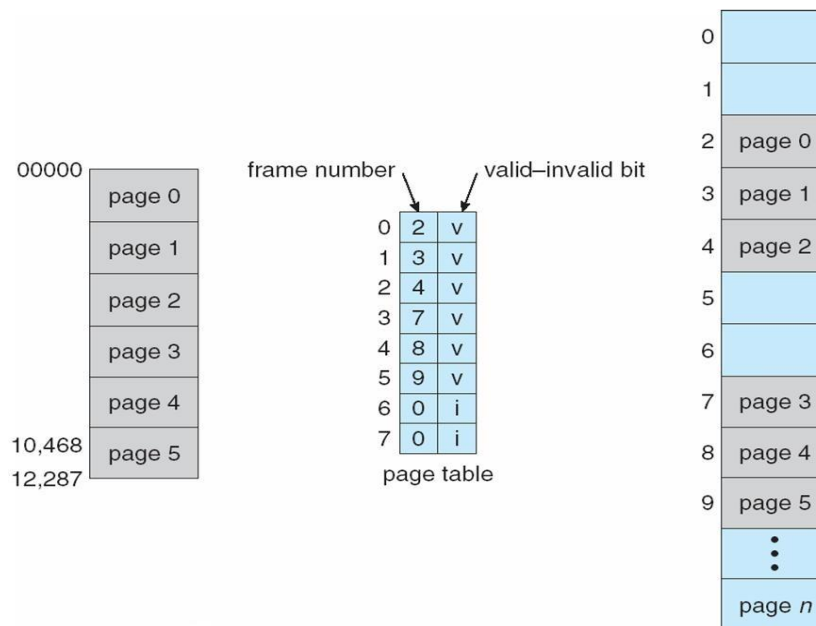
## Effective Access Time

- Associative Lookup = e time unit
- Assume memory cycle time is 1 microsecond
- Hit ratio – percentage of times that a page number is found in the associative registers; ratio related to number of associative registers
- Hit ratio = an **Effective Access Time (EAT)**

$$\begin{aligned} \text{EAT} &= (1 + e) a + (2 + e)(1 - a) \\ &= 2 + e - a \end{aligned}$$

## Memory Protection

- Memory protection implemented by associating protection bit with each frame
- **Valid-invalid** bit attached to each entry in the page table:
- “valid” indicates that the associated page is in the process’ logical address space, and is thus a legal page
- “invalid” indicates that the page is not in the process’ logical address space
- **Valid (v) or Invalid (i) Bit In A Page Table**



## Shared Pages

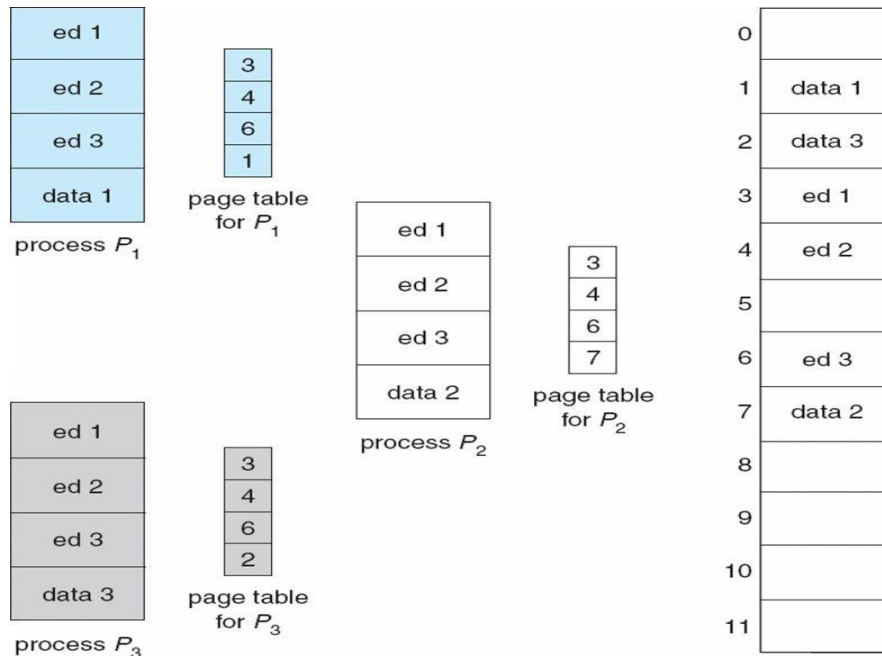
### Shared code

- One copy of read-only (reentrant) code shared among processes (i.e., text editors, compilers, window systems).
- Shared code must appear in same location in the logical address space of all processes

### Private code and data

- Each process keeps a separate copy of the code and data
- The pages for the private code and data can appear anywhere in the logical address space

## Shared Pages Example



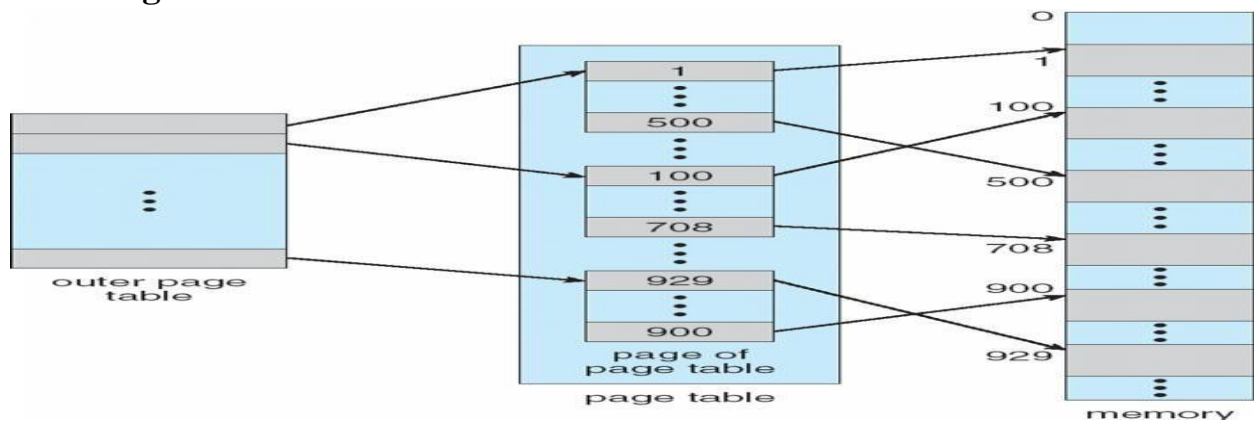
## Structure of the Page Table

- Hierarchical Paging
- Hashed Page Tables
- Inverted Page Tables

## Hierarchical Page Tables

- Break up the logical address space into multiple page tables
- A simple technique is a two-level page table

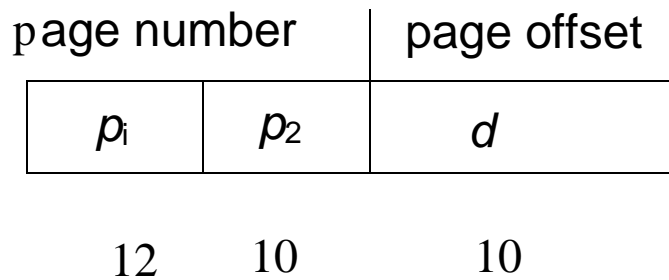
## Two-Level Page-Table Scheme



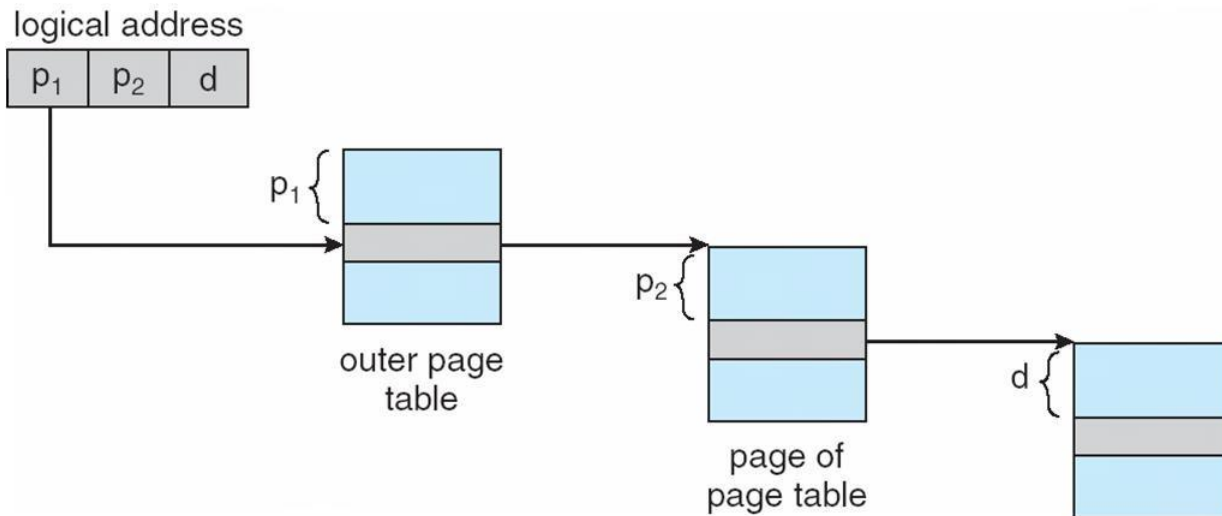
## Two-Level Paging Example

- A logical address (on 32-bit machine with 1K page size) is divided into:
  - a page number consisting of 22 bits
  - a page offset consisting of 10 bits
- Since the page table is paged, the page number is further divided into:
  - a 12-bit page number
  - a 10-bit page offset
- Thus, a logical address is as follows:

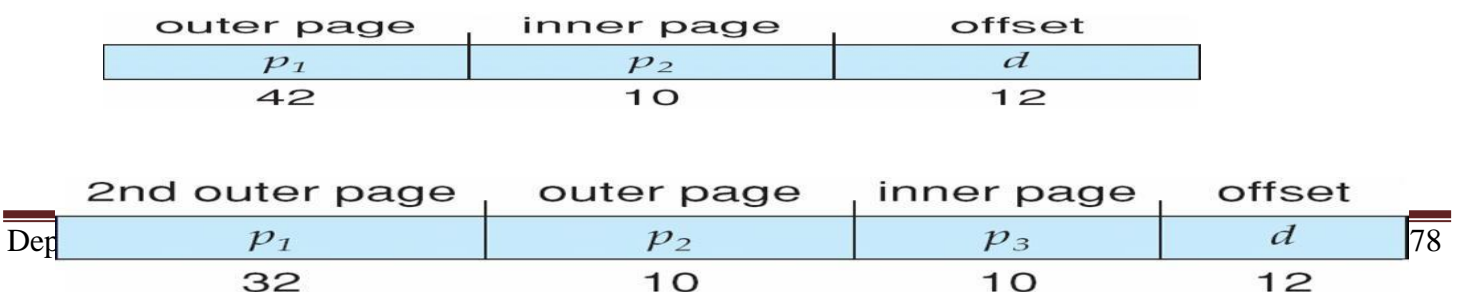
where  $p_i$  is an index into the outer page table, and  $p_2$  is the displacement within the page of the outer page table



## Address-Translation Scheme



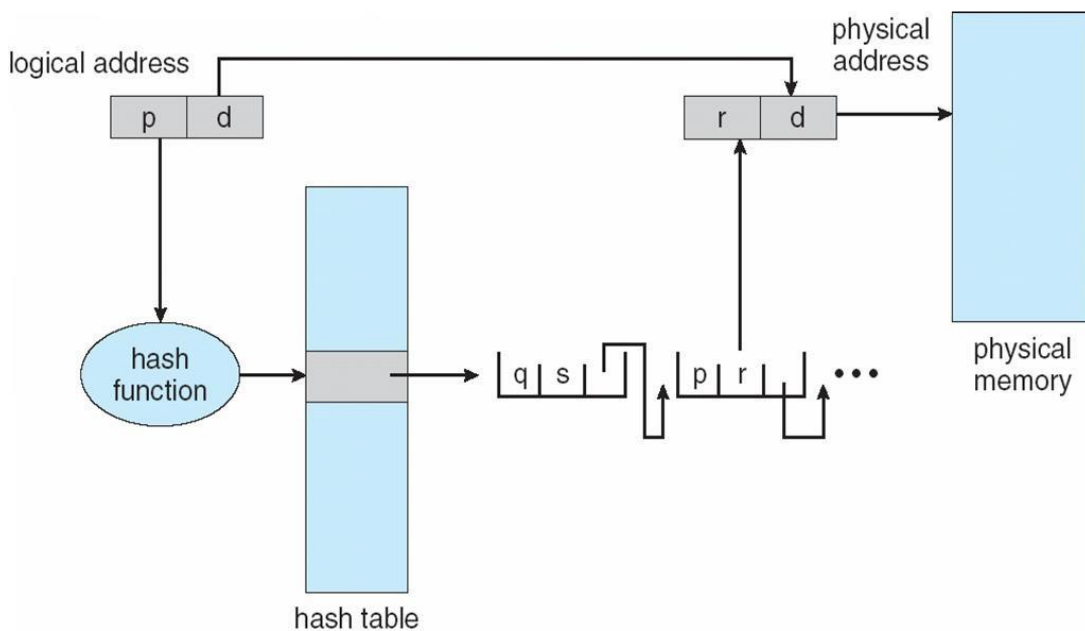
## Three-level Paging Scheme



## Hashed Page Tables

- Common in address spaces  $> 32$  bits
- The virtual page number is hashed into a page table
- This page table contains a chain of elements hashing to the same location
- Virtual page numbers are compared in this chain searching for a match
- If a match is found, the corresponding physical frame is extracted

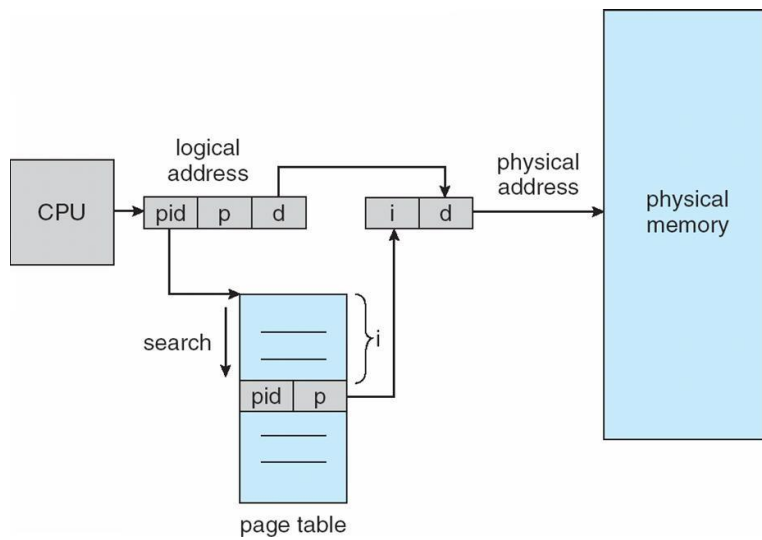
## Hashed Page Table



## Inverted Page Table

- One entry for each real page of memory
  - Entry consists of the virtual address of the page stored in that real memory location, with information about the process that owns that page
  - Decreases memory needed to store each page table, but increases time needed to search the table when a page reference occurs
  - Use hash table to limit the search to one — or at most a few — page-table entries
-

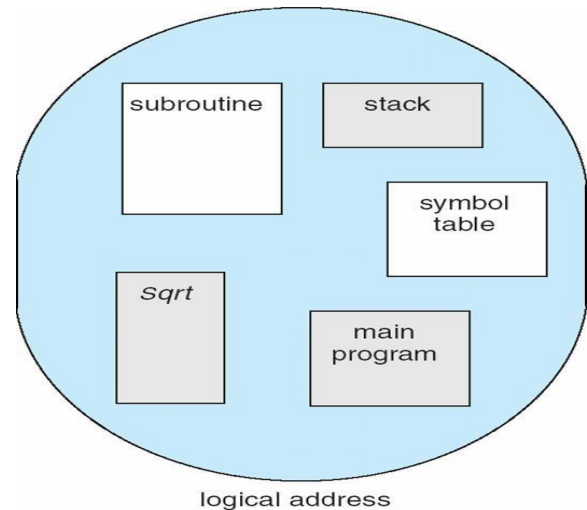
## Inverted Page Table Architecture



## Segmentation

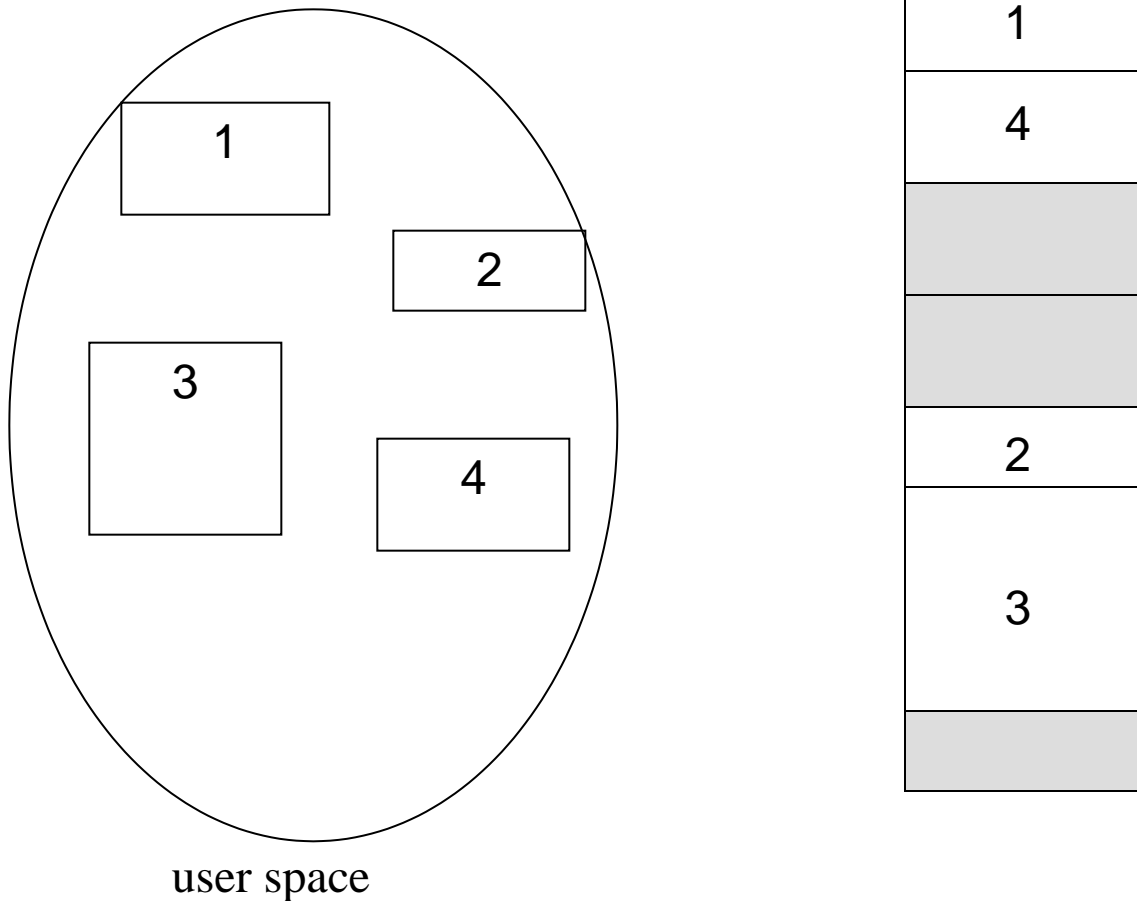
Memory-management scheme that supports user view of memory

- A program is a collection of segments
- A segment is a logical unit such as:
  - main program
  - procedure function
  - method
  - object
  - local variables, global variables
  - common block
  - stack
  - symbol table
  - arrays



## User's View of a Program

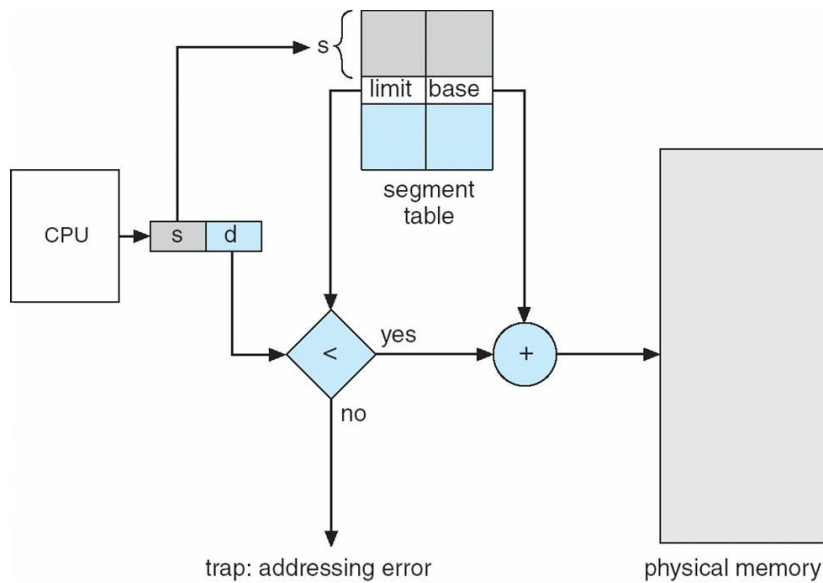
## Logical View of Segmentation



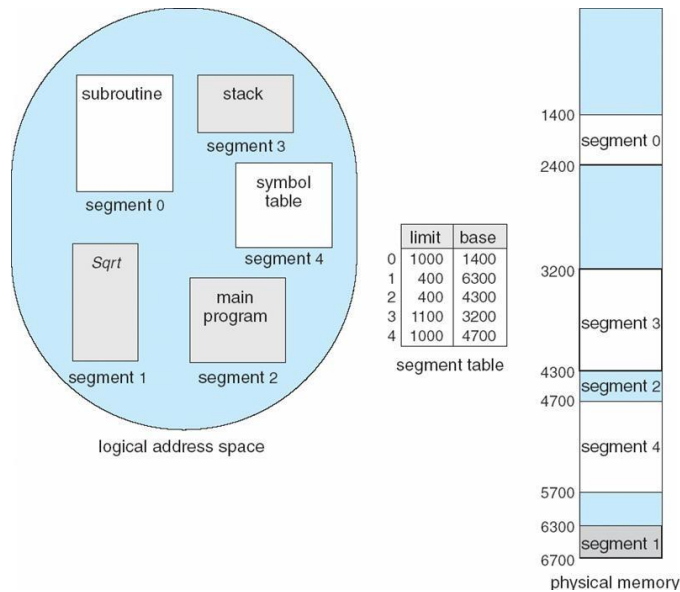
## Segmentation Architecture

- Logical address consists of a two tuple:
  - $\langle \text{segment-number}, \text{offset} \rangle$ ,
- **Segment table** – maps two-dimensional physical addresses, each to a memory location in physical memory space
- **base** – contains the starting physical address where the segments reside in memory
- **limit** – specifies the length of the segment
- **Segment-table base register (STBR)** points to the segment table's location in memory
- **Segment-table length register (STLR)** indicates number of segments used by a program; segment number  $s$  is legal if  $s < \text{STLR}$
- Protection
- With each entry in segment table associate:
  - $\square$  validation bit = 0  $\text{P}$  illegal segment
  - $\square$  read/write/execute privileges
- Protection bits associated with segments; code sharing occurs at segment level
- Since segments vary in length, memory allocation is a dynamic storage-allocation problem
- A segmentation example is shown in the following diagram

## Segmentation Hardware



## Example of Segmentation

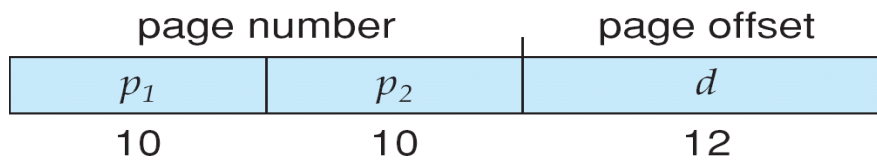
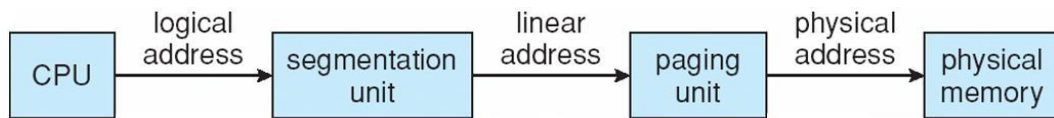


## Example: The Intel Pentium

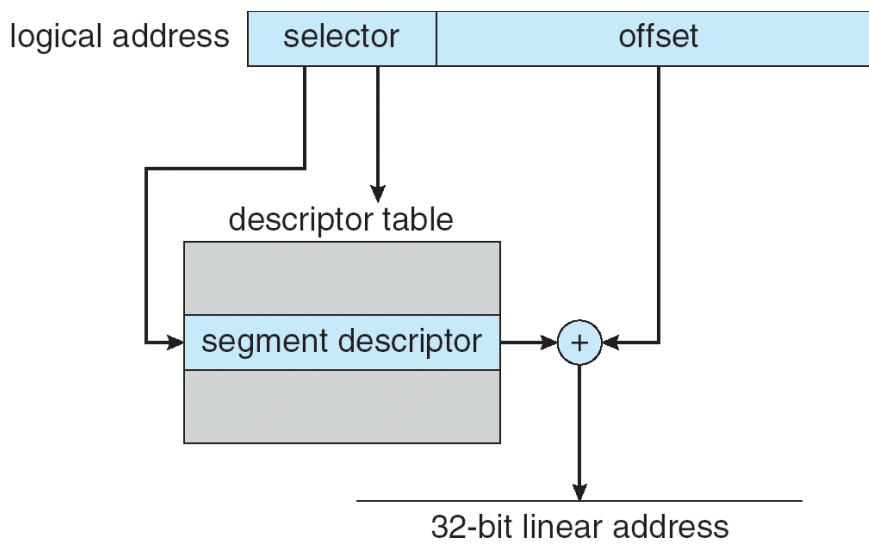
- Supports both segmentation and segmentation with paging
- CPU generates logical address
- Given to segmentation unit
  - Which produces linear addresses
- Linear address given to paging unit
  - Which generates physical address in main memory
  - Paging units form equivalent of MMU



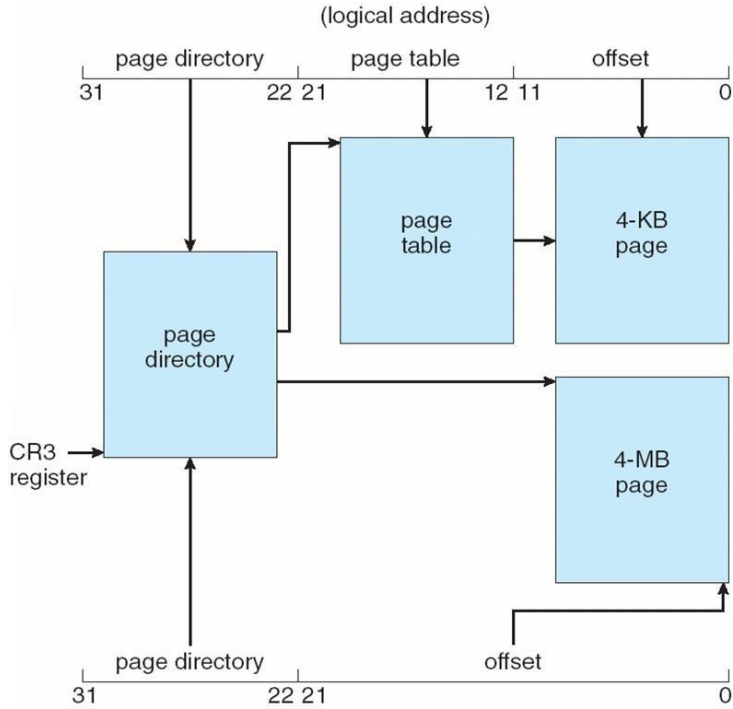
## Logical to Physical Address Translation in Pentium



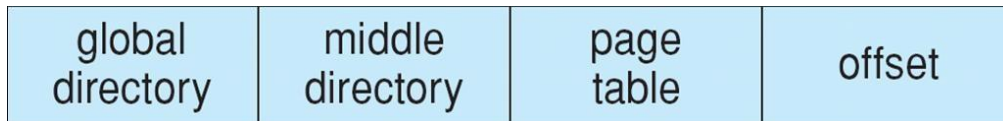
## Intel Pentium Segmentation



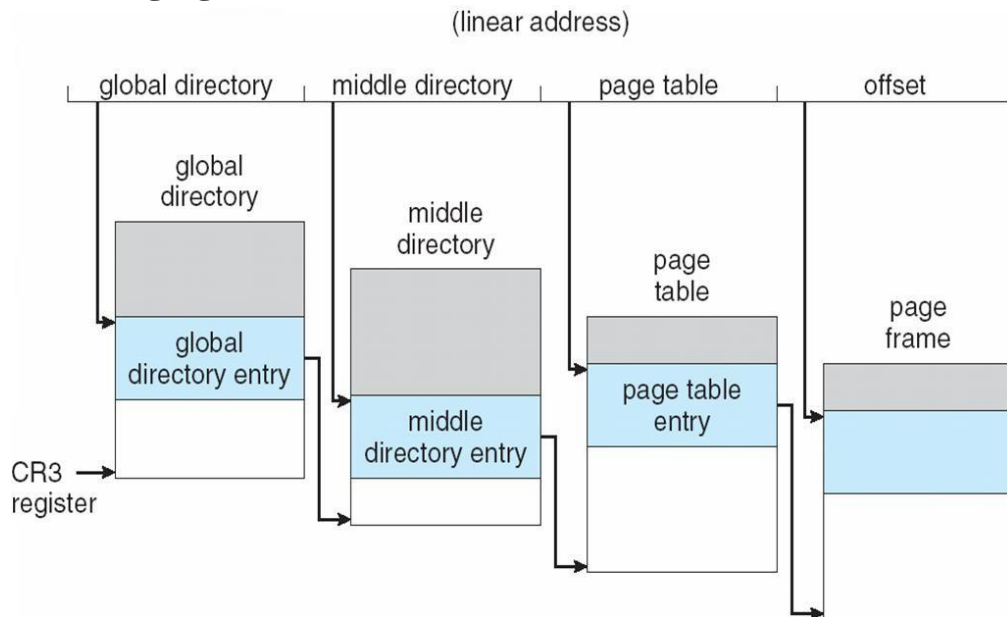
# Pentium Paging Architecture

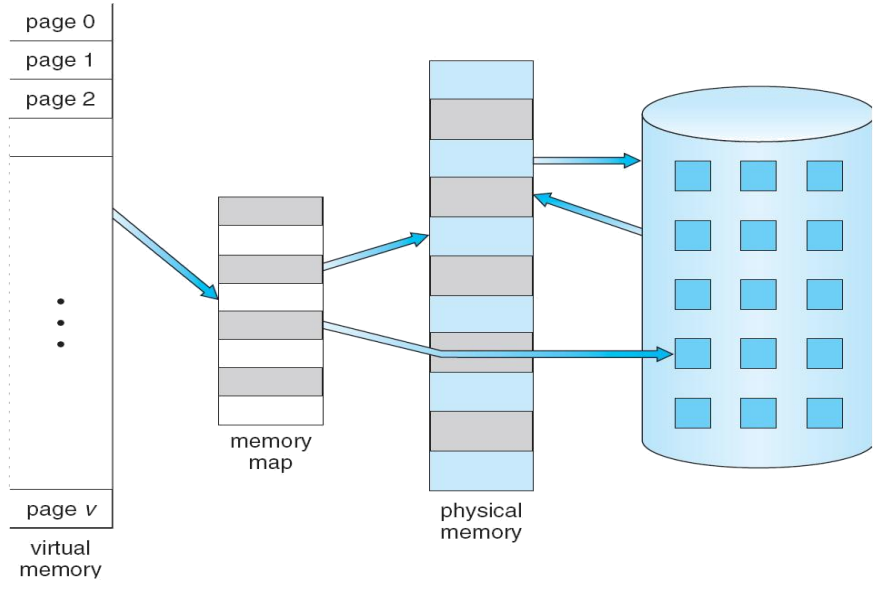


## Linear Address in Linux



## Three-level Paging in Linux





# UNIT -5



# VIRTUAL MEMORY

## Objective

- To describe the benefits of a virtual memory system.
- To explain the concepts of demand paging, page-replacement algorithms, and allocation of page frames.
- To discuss the principle of the working-set model.

## Virtual Memory

- Virtual memory is a technique that allows the execution of process that may not be completely in memory. The main visible advantage of this scheme is that programs can be larger than physical memory.
  - Virtual memory is the separation of user logical memory from physical memory this separation allows an extremely large virtual memory to be provided for programmers when only a smaller physical memory is available ( Fig ).
  - Following are the situations, when entire program is not required to load fully.
    1. User written error handling routines are used only when an error occurs in the data or computation.
    2. Certain options and features of a program may be used rarely.
    3. Many tables are assigned a fixed amount of address space even though only a small amount of the table is actually used.
  - The ability to execute a program that is only partially in memory would counter many benefits.
    1. Less number of I/O would be needed to load or swap each user program into memory.
    2. A program would no longer be constrained by the amount of physical memory that is available.
    3. Each user program could take less physical memory, more programs could be run the same time, with a corresponding increase in CPU utilization and throughput.
-

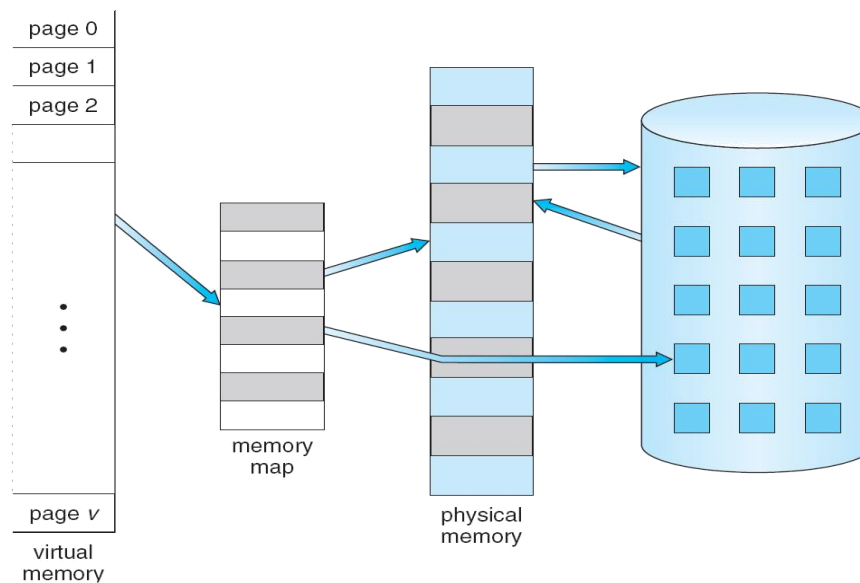


Fig. Diagram showing virtual memory that is larger than physical memory.

Virtual memory is commonly implemented by demand paging. It can also be implemented in a segmentation system. Demand segmentation can also be used to provide virtual memory.

## Demand Paging

A demand paging is similar to a paging system with swapping(Fig 5.2). When we want to execute a process, we swap it into memory. Rather than swapping the entire process into memory.

When a process is to be swapped in, the pager guesses which pages will be used before the process is swapped out again. Instead of swapping in a whole process, the pager brings only those necessary pages into memory. Thus, it avoids reading into memory pages that will not be used in anyway, decreasing the swap time and the amount of physical memory needed.

Hardware support is required to distinguish between those pages that are in memory and those pages that are on the disk using the valid-invalid bit scheme. Where valid and invalid pages can be checked checking the bit and marking a page will have no effect if the process never attempts to access the pages. While the process executes and accesses pages that are memory resident, execution proceeds normally.

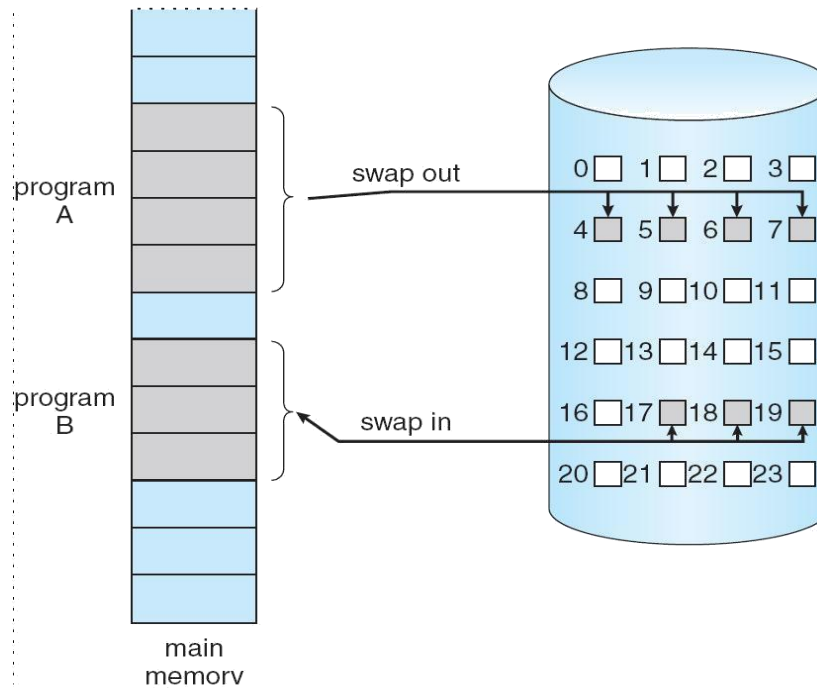


Fig. Transfer of a paged memory to continuous disk space

Access to a page marked invalid causes a page-fault trap. This trap is the result of the operating system's failure to bring the desired page into memory. But page fault can be handled as following (Fig 5.3):

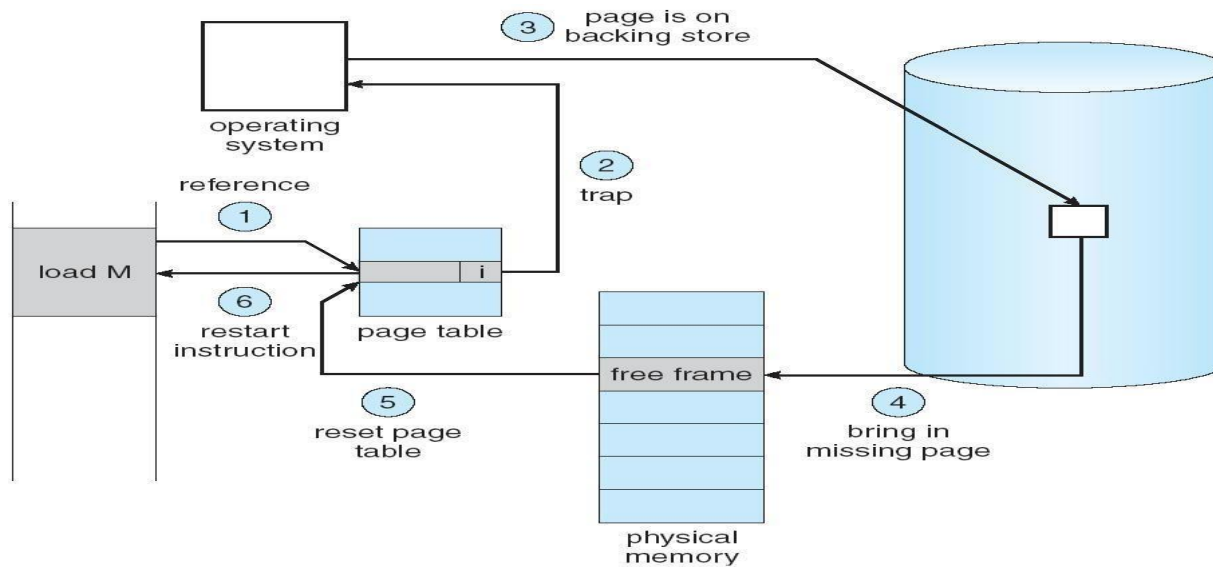


Fig. Steps in handling a page fault

1. We check an internal table for this process to determine whether the reference was a valid or invalid memory access.
2. If the reference was invalid, we terminate the process. If it was valid, but we have not yet brought in that page, we now page in the latter.
3. We find a free frame.
4. We schedule a disk operation to read the desired page into the newly allocated frame.
5. When the disk read is complete, we modify the internal table kept with the process and the page table to indicate that the page is now in memory.
6. We restart the instruction that was interrupted by the illegal address trap. The process can now access the page as though it had always been memory.

Therefore, the operating system reads the desired page into memory and restarts the process as though the page had always been in memory.

The page replacement is used to make the frame free if they are not in used. If no frame is free then other process is called in.

### **Advantages of Demand Paging:**

1. Large virtual memory.
2. More efficient use of memory.
3. Unconstrained multiprogramming. There is no limit on degree of multiprogramming.

### **Disadvantages of Demand Paging:**

4. Number of tables and amount of processor overhead for handling page interrupts are greater than in the case of the simple paged management techniques.
5. due to the lack of an explicit constraints on a jobs address space size.

### **Page Replacement Algorithm**

There are many different page replacement algorithms. We evaluate an algorithm by running it on a particular string of memory reference and computing the number of page faults. The string of memory references is called reference string. Reference strings are generated artificially or by tracing a given system and recording the address of each memory reference. The latter choice produces a large number of data.

1. For a given page size we need to consider only the page number, not the entire address.
2. if we have a reference to a page p, then any immediately following references to page p will never cause a page fault. Page p will be in memory after the first reference; the immediately following references will not fault.

Eg:- consider the address sequence

0100, 0432, 0101, 0612, 0102, 0103, 0104, 0101, 0611, 0102, 0103, 0104, 0101, 0610, 0102,

---



0103, 0104, 0104, 0101, 0609, 0102, 0105  
and reduce to 1, 4, 1, 6, 1, 6, 1, 6, 1, 6, 1

To determine the number of page faults for a particular reference string and page replacement algorithm, we also need to know the number of page frames available. As the number of frames available increase, the number of page faults will decrease.

## **FIFO Algorithm**

The simplest page-replacement algorithm is a FIFO algorithm. A FIFO replacement algorithm associates with each page the time when that page was brought into memory. When a page must be replaced, the oldest page is chosen. We can create a FIFO queue to hold all pages in memory.

The first three references (7, 0, 1) cause page faults, and are brought into these empty eg. 7, 0, 1, 2, 0, 3, 0, 4, 2, 3, 0, 3, 2, 1, 2, 0, 1 and consider 3 frames. This replacement means that the next reference to 0 will fault. Page 1 is then replaced by page 0.

## **Optimal Algorithm**

An optimal page-replacement algorithm has the lowest page-fault rate of all algorithms. An optimal page-replacement algorithm exists, and has been called OPT or MIN. It is simply

Replace the page that will not be used for the longest period of time.

Now consider the same string with 3 empty frames.

The reference to page 2 replaces page 7, because 7 will not be used until reference 15, whereas page 0 will be used at 5, and page 1 at 14. The reference to page 3 replaces page 1, as page 1 will be the last of the three pages in memory to be referenced again. Optimal replacement is much better than a FIFO.

The optimal page-replacement algorithm is difficult to implement, because it requires future knowledge of the reference string.

## **LRU Algorithm**

The FIFO algorithm uses the time when a page was brought into memory; the OPT algorithm uses the time when a page is to be used. In LRU replace the page that has not been used for the longest period of time.

LRU replacement associates with each page the time of that page's last use. When a page must be replaced, LRU chooses that page that has not been used for the longest period of time.

Let  $S^R$  be the reverse of a reference string  $S$ , then the page-fault rate for the OPT algorithm on  $S$  is the same as the page-fault rate for the OPT algorithm on  $S^R$ .

---

## LRU Approximation Algorithms

Some systems provide no hardware support, and other page-replacement algorithm. Many systems provide some help, however, in the form of a reference bit. The reference bit for a page is set, by the hardware, whenever that page is referenced. Reference bits are associated with each entry in the page table. Initially, all bits are cleared (to 0) by the operating system. As a user process executes, the bit associated with each page referenced is set (to 1) by the hardware.

### Additional-Reference-Bits Algorithm

The operating system shifts the reference bit for each page into the high-order or of its 5-bit byte, shifting the other bits right 1 bit, discarding the low-order bit.

These 5-bit shift registers contain the history of page use for the last eight time periods. If the shift register contains 00000000, then the page has not been

used for eight time periods; a page that is used at least once each period would have a shift register value of 11111111.

### Second-Chance Algorithm

The basic algorithm of second-chance replacement is a FIFO replacement algorithm. When a page gets a second chance, its reference bit is cleared and its arrival time is reset to the current time.

### Enhanced Second-Chance Algorithm

The second-chance algorithm described above can be enhanced by considering both the reference bit and the modify bit as an ordered pair.

1. (0,0) neither recently used nor modified best page to replace.
2. (0,1) not recently used but modified not quite as good, because the page will need to be written out before replacement.
3. (1,0) recently used but clean probably will be used again soon.
4. (1,1) recently used and modified probably will be used again, and write out will be needed before replacing it

### Counting Algorithms

There are many other algorithms that can be used for page replacement.

- **LFU Algorithm:** The least frequently used (LFU) page-replacement algorithm requires that the page with the smallest count be replaced. This algorithm suffers from the situation in which a page is used heavily during the initial phase of a process, but then is never used again.
-

- **MFU Algorithm:** The most frequently used (MFU) page-replacement algorithm is based on the argument that the page with the smallest count was probably just brought in and has yet to be used.

### **Page Buffering Algorithm**

When a page fault occurs, a victim frame is chosen as before. However, the desired page is read into a free frame from the pool before the victim is written out.

This procedure allows the process to restart as soon as possible, without waiting for the victim page to be written out. When the victim is later written out, its frame is added to the free-frame pool.

When the FIFO replacement algorithm mistakenly replaces a page mistakenly replaces a page that is still in active use, that page is quickly retrieved from the free-frame buffer, and no I/O is necessary. The free-frame buffer provides protection against the relatively poor, but simple, FIFO replacement algorithm.



# UNIT -6



## Principles of deadlock

To develop a description of deadlocks, which prevent sets of concurrent processes from completing their tasks. To present a number of different methods for preventing or avoiding deadlocks in a computer system

### **The Deadlock Problem**

A set of blocked processes each holding a resource and waiting to acquire a resource held by another process in the set

Example

System has 2 disk drives

$P_1$  and  $P_2$  each hold one disk drive and each needs another one

Example

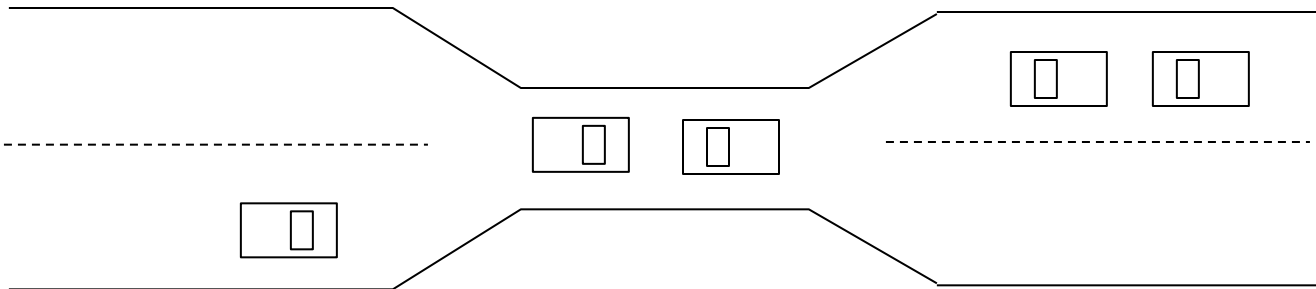
semaphores  $A$  and  $B$ , initialized to 1

$P_0$                        $P_1$

wait (A);              wait(B)

wait (B);              wait(A)

### **Bridge Crossing Example**



Traffic only in one direction

Each section of a bridge can be viewed as a resource

If a deadlock occurs, it can be resolved if one car backs up (preempt resources and rollback)

Several cars may have to be backed up if a deadlock occurs

Starvation is possible

Note – Most OSes do not prevent or deal with deadlocks

## System Model

Resource types  $R_1, R_2, \dots, R_m$

*CPU cycles, memory space, I/O devices*

Each resource type  $R_i$  has  $W_i$  instances.

Each process utilizes a resource as follows:

**request**

**use**

**release**

## Deadlock Characterization

Deadlock can arise if four conditions hold simultaneously

**Mutual exclusion:** only one process at a time can use a resource

**Hold and wait:** a process holding at least one resource is waiting to acquire additional resources held by other processes

**No preemption:** a resource can be released only voluntarily by the process holding it, after that process has completed its task

**Circular wait:** there exists a set  $\{P_0, P_1, \dots, P_n\}$  of waiting processes such that  $P_0$  is waiting for a resource that is held by  $P_1$ ,  $P_1$  is waiting for a resource that is held by

$P_2, \dots, P_{n-1}$  is waiting for a resource that is held by

$P_n$ , and  $P_0$  is waiting for a resource that is held by  $P_0$ .

n

## Resource-Allocation Graph

A set of vertices  $V$  and a set of edges  $E$

$V$  is partitioned into two types:

$P = \{P_1, P_2, \dots, P_n\}$ , the set consisting of all the processes in the system

$R = \{R_1, R_2, \dots, R_m\}$ , the set consisting of all resource types in the system

request edge – directed edge  $P_i \rightarrow R_j$

assignment edge – directed edge  $R_j \rightarrow P_i$

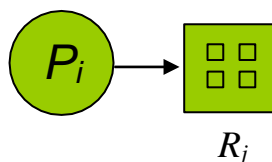
Process



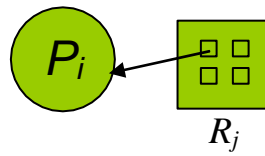
Resource Type with 4 instances



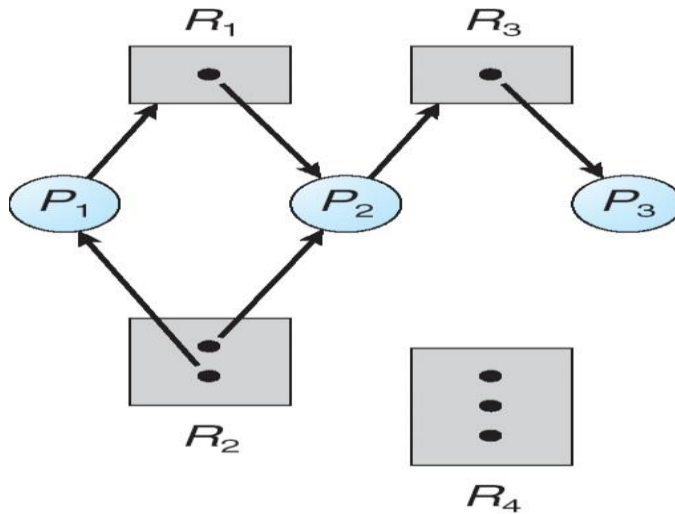
$P_i$  requests instance of  $R_j$



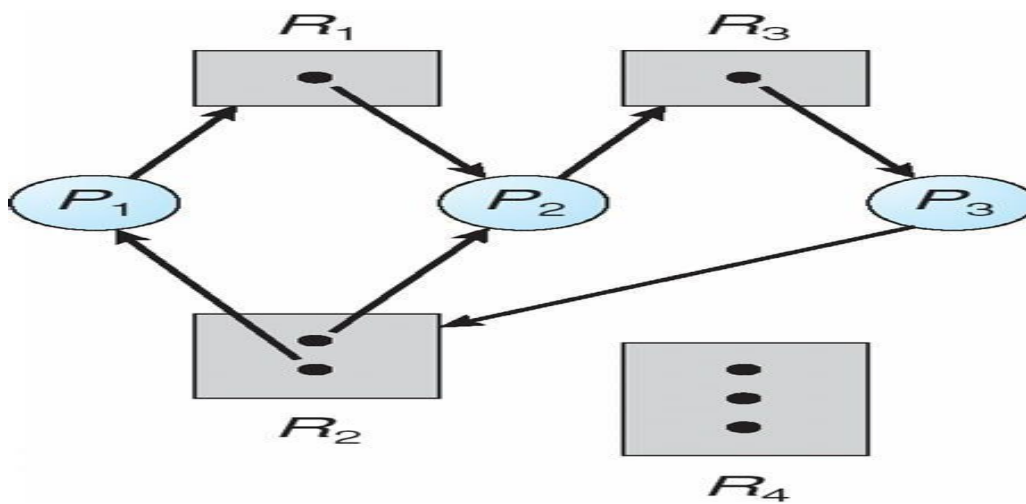
$P_i$  is holding an instance of  $R_j$



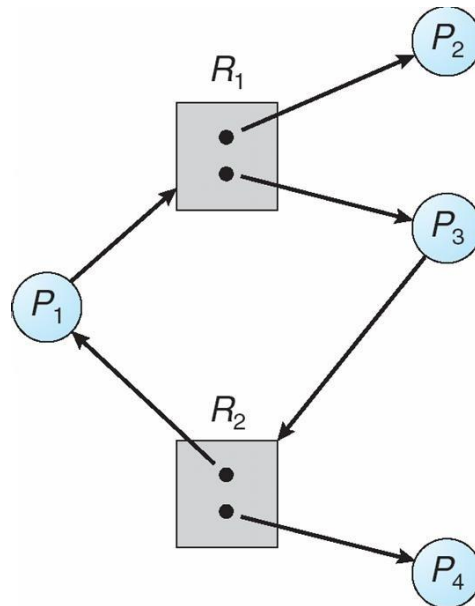
### Example of a Resource Allocation Graph



### Resource Allocation Graph With A Deadlock



## Graph With A Cycle But No Deadlock



### Basic Facts

If graph contains no cycles ⇒ no deadlock  
If graph contains a cycle ⇒ if only one instance per resource type, then deadlock

if several instances per resource type, possibility of deadlock

### Methods for Handling Deadlocks

Ensure that the system will *never* enter a deadlock state  
Allow the system to enter a deadlock state and then recover  
Ignore the problem and pretend that deadlocks never occur in the system; used by most operating systems, including UNIX

### Deadlock Prevention

Restrain the ways request can be made

**Mutual Exclusion** – not required for sharable resources; must hold for nonsharable resources

**Hold and Wait** – must guarantee that whenever a process requests a resource, it does not hold any other resources

Require process to request and be allocated all its resources before it begins execution, or allow process to request resources only when the process has none

Low resource utilization; starvation possible

**No Preemption** –

If a process that is holding some resources requests another resource that cannot be immediately allocated to it, then all resources currently being held are released

Preempted resources are added to the list of resources for which the process is waiting

Process will be restarted only when it can regain its old resources, as well as the new ones that it is requesting

**Circular Wait** – impose a total ordering of all resource types, and require that each process requests resources in an increasing order of enumeration



## Deadlock Avoidance

Requires that the system has some additional *a priori* information available

Simplest and most useful model requires that each process declare the *maximum number* of resources of each type that it may need

The deadlock-avoidance algorithm dynamically examines the resource-allocation state to ensure that there can never be a circular-wait condition

Resource-allocation *state* is defined by the number of available and allocated resources, and the maximum demands of the processes

### Safe State

When a process requests an available resource, system must decide if immediate allocation leaves the system in a safe state

System is in safe state if there exists a sequence  $\langle P_1, P_2, \dots, P_n \rangle$  of ALL the processes in the system such that for each  $P_i$ , the resources that  $P_i$  can still request can be satisfied by currently available resources + resources held by all the  $P_j$ , with  $j < i$ . That is:

If  $P_i$  resource needs are not immediately available, then  $P_i$  can wait until all  $P_j$  have finished

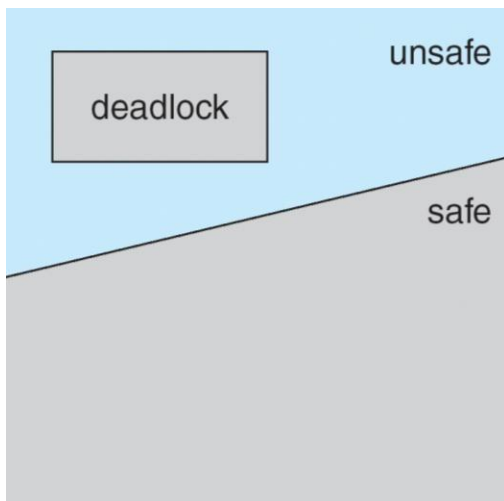
When  $P_j$  is finished,  $P_i$  can obtain needed resources, execute, return allocated resources, and terminate

When  $P_i$  terminates,  $P_{i+1}$  can obtain its needed resources, and so on

### Basic Facts

If a system is in safe state  $\nexists$  no deadlocks  
If a system is in unsafe state  $\nexists$  possibility of deadlock  
Avoidance  $\nexists$  ensure that a system will never enter an unsafe state.

### Safe, Unsafe, Deadlock State



### Avoidance algorithms

Single instance of a resource type

Use a resource-allocation graph

Multiple instances of a resource type

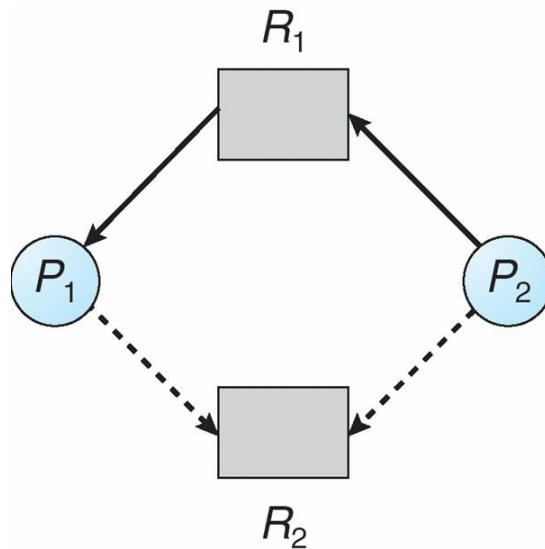
Use the banker's algorithm

## Resource-Allocation Graph Scheme

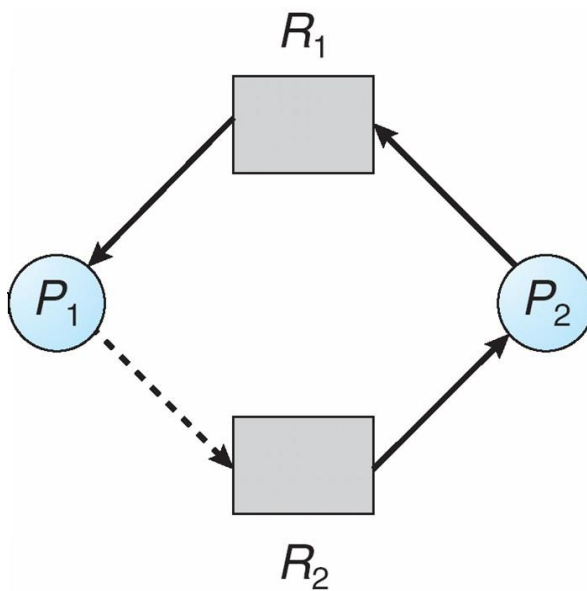
nClaim edge  $P_i \rightarrow R_j$  indicated that process  $P_i$  may request resource  $R_j$ ; represented by a dashed line  
nClaim edge converts to request edge when a process requests a resource  
nRequest edge converted to an assignment edge when the resource is allocated to the process

nWhen a resource is released by a process, assignment edge reconverts to a claim edge  
nResources must be claimed *a priori* in the system

## Resource-Allocation Graph



## Unsafe State In Resource-Allocation Graph



## Resource-Allocation Graph Algorithm

Suppose that process  $P_i$  requests a resource  $R_j$ .

The request can be granted only if converting the request edge to an assignment edge does not result in the formation of a cycle in the resource allocation graph

## Banker's Algorithm

Multiple instances  
Each process must a priori claim maximum use  
When a process requests a resource it may have to wait  
When a process gets all its resources it must return them in a finite amount of time

## Data Structures for the Banker's Algorithm

Let  $n$  = number of processes, and  $m$  = number of resources types.

**Available:** Vector of length  $m$ . If available  $[j] = k$ , there are  $k$  instances of resource type  $R_j$  available

**Max:**  $n \times m$  matrix. If  $Max [i,j] = k$ , then process  $P_i$  may request at most  $k$  instances of resource type  $R_j$

**Allocation:**  $n \times m$  matrix. If  $Allocation[i,j] = k$  then  $P_i$  is currently allocated  $k$  instances of  $R_j$   
**Need:**  $n \times m$  matrix. If  $Need[i,j] = k$ , then  $P_i$  may need  $k$  more instances of  $R_j$  to complete its task

$$Need [i,j] = Max[i,j] - Allocation [i,j]$$

## Safety Algorithm

1. Let *Work* and *Finish* be vectors of length  $m$  and  $n$ , respectively. Initialize:

*Work* = *Available*

*Finish*  $[i] = false$  for  $i = 0, 1, \dots, n-1$

2. Find and  $i$  such that both:

(a) *Finish*  $[i] = false$  (b)  $Need_i \leq Work$

If no such  $i$  exists, go to step 4

3. *Work* = *Work* + *Allocation* <sub>$i$</sub>

*Finish* $[i] = true$

go to step 2

4. If *Finish*  $[i] == true$  for all  $i$ , then the system is in a safe state

## Resource-Request Algorithm for Process $P_i$

*Request* = request vector for process  $P_i$ . If  $Request_i [j] = k$  then process  $P_i$  wants  $k$  instances of resource type  $R_j$ .

1. If  $Request_i \leq Need_i$  go to step 2. Otherwise, raise error condition, since process has exceeded its maximum claim

2. If  $Request_i \leq Available$ , go to step 3. Otherwise  $P_i$  must wait, since resources are not available

3. Pretend to allocate requested resources to  $P_i$  by modifying the state as follows:

*Available* = *Available* - *Request*;

*Allocation* <sub>$i$</sub>  = *Allocation* <sub>$i$</sub>  + *Request*;

*Need* <sub>$i$</sub>  = *Need* <sub>$i$</sub>  - *Request*;

If safe  $P$  the resources are allocated to  $P_i$

If unsafe  $P$   $P_i$  must wait, and the old resource-allocation state is restored

---

## Example of Banker's Algorithm

5 processes  $P_0$  through  $P_4$ ;

3 resource types:

A (10 instances), B (5 instances), and C (7 instances)

Snapshot at time  $T_0$ :

|   |   |   | <u>Allocation</u> |   |   | <u>Max Available</u> |   |   |   |
|---|---|---|-------------------|---|---|----------------------|---|---|---|
| A | B | C | A                 | B | C | A                    | B | C |   |
|   |   |   | $P_0$             | 0 | 1 | 0                    | 7 | 5 | 3 |
|   |   |   | $P_1$             | 2 | 0 | 0                    | 3 | 2 | 2 |
|   |   |   | $P_2$             | 3 | 0 | 2                    | 9 | 0 | 2 |
|   |   |   | $P_3$             | 2 | 1 | 1                    | 2 | 2 | 2 |
|   |   |   | $P_4$             | 0 | 0 | 2                    | 4 | 3 | 3 |

The content of the matrix *Need* is defined to be  $Max - Allocation$

|   |   |   | <u>Need</u> |   |   |   |
|---|---|---|-------------|---|---|---|
| A | B | C | A           | B | C |   |
|   |   |   | $P_0$       | 7 | 4 | 3 |
|   |   |   | $P_1$       | 1 | 2 | 2 |
|   |   |   | $P_2$       | 6 | 0 | 0 |
|   |   |   | $P_3$       | 0 | 1 | 1 |
|   |   |   | $P_4$       | 4 | 3 | 1 |

The system is in a safe state since the sequence  $\langle P_1, P_3, P_4, P_2, P_0 \rangle$  satisfies safety criteria

### Example: $P_1$ Request (1,0,2)

Check that Request  $\leq$  Available (that is, (1,0,2)  $\leq$  (3,3,2))  $\wedge$  true

|   |   |   | <u>Allocation</u> |   |   | <u>Need</u> |   |   |   |
|---|---|---|-------------------|---|---|-------------|---|---|---|
| A | B | C | A                 | B | C | A           | B | C |   |
|   |   |   | $P_0$             | 0 | 1 | 0           | 7 | 4 | 3 |
|   |   |   | $P_1$             | 3 | 0 | 2           | 0 | 2 | 0 |
|   |   |   | $P_2$             | 3 | 0 | 1           | 6 | 0 | 0 |
|   |   |   | $P_3$             | 2 | 1 | 1           | 0 | 1 | 1 |
|   |   |   | $P_4$             | 0 | 0 | 2           | 4 | 3 | 1 |

Executing safety algorithm shows that sequence  $\langle P_1, P_3, P_4, P_0, P_2 \rangle$  satisfies safety requirement

Can request for (3,3,0) by  $P_4$  be granted?

Can request for (0,2,0) by  $P_0$  be granted?

## Deadlock Detection

Allow system to enter deadlock state Detection algorithm Recovery scheme

## Single Instance of Each Resource Type

Maintain *wait-for* graph

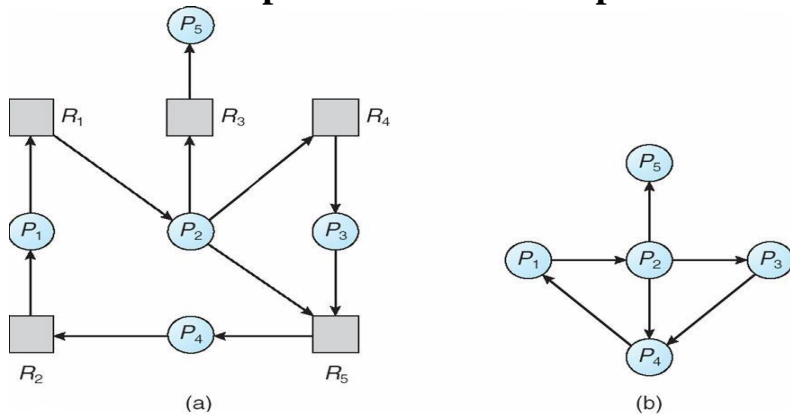
Nodes are processes

$P_i \text{ @ } P_j$  if  $P_i$  is waiting for  $P_j$

Periodically invoke an algorithm that searches for a cycle in the graph. If there is a cycle, there exists a deadlock

An algorithm to detect a cycle in a graph requires an order of  $n^2$  operations, where  $n$  is the number of vertices in the graph

## Resource-Allocation Graph and Wait-for Graph



Resource-Allocation Graph

Corresponding wait-for graph

## Several Instances of a Resource Type

**Available:** A vector of length  $m$  indicates the number of available resources of each type. **Allocation:** An  $n \times m$  matrix defines the number of resources of each type currently allocated to each process. **Request:** An  $n \times m$  matrix indicates the current request of each process. If  $Request[i][j] = k$ , then process  $P_i$  is requesting  $k$  more instances of resource type  $R_j$ .

## Detection Algorithm

- Let *Work* and *Finish* be vectors of length  $m$  and  $n$ , respectively Initialize:
    - $Work = Available$
    - For  $i = 1, 2, \dots, n$ , if  $Allocation_i \neq 0$ , then  $Finish[i] = false$ ; otherwise,  $Finish[i] = true$ . Find an index  $i$  such that both:
      - $Finish[i] == false$
      - $Request_i \leq Work$
 If no such  $i$  exists, go to step 4
  - $Work = Work + Allocation_i$   
 $Finish[i] = true$
- go to step 2. If  $Finish[i] == false$ , for some  $i$ ,  $1 \leq i \leq n$ , then the system is in deadlock state. Moreover, if  $Finish[i] == false$ , then  $P_i$  is deadlocked

**Algorithm requires an order of  $O(m \times n^2)$  operations to detect whether the system is in deadlocked state**

## Example of Detection Algorithm

Five processes  $P_0$  through  $P_4$ ; three resource types  
 A (7 instances), B (2 instances), and C (6 instances)  
 Snapshot at time  $T_0$ :

|       | <u>Allocation</u> | <u>Request</u> | <u>Available</u> |
|-------|-------------------|----------------|------------------|
|       | A B C             | A B C          | A B C            |
| $P_0$ | 0 1 0             | 0 0 0          | 0 0 0            |
| $P_1$ | 2 0 0             | 2 0 2          |                  |
| $P_2$ | 3 0 3             | 0 0 0          |                  |
| $P_3$ | 2 1 1             | 1 0 0          |                  |
| $P_4$ | 0 0 2             | 0 0 2          |                  |

Sequence  $\langle P_0, P_2, P_3, P_1, P_4 \rangle$  will result in  $Finish[i] = \text{true}$  for all  $i$

$P_2$  requests an additional instance of type C

|       | <u>Request</u> |
|-------|----------------|
|       | A B C          |
| $P_0$ | 0 0 0          |
| $P_1$ | 2 0 1          |
| $P_2$ | 0 0 1          |
| $P_3$ | 1 0 0          |
| $P_4$ | 0 0 2          |

State of system?

Can reclaim resources held by process  $P_0$ , but insufficient resources to fulfill other processes; requests  
 Deadlock exists, consisting of processes  $P_1, P_2, P_3$ , and  $P_4$

## Detection-Algorithm Usage

When, and how often, to invoke depends on:

How often a deadlock is likely to occur?

How many processes will need to be rolled back?

□one for each disjoint cycle  
 If detection algorithm is invoked arbitrarily, there may be many cycles in the resource graph and so we would not be able to tell which of the many deadlocked processes “caused” the deadlock

## Recovery from Deadlock: Process Termination

Abort all deadlocked processes  
 Abort one process at a time until the deadlock cycle is eliminated  
 In which order should we choose to abort?

Priority of the process

How long process has computed, and how much longer to completion

Resources the process has used

Resources process needs to complete

How many processes will need to be terminated

Is process interactive or batch?