

Operating Systems – 1BCS304

Module -V

Course Coordinator:
Mrs. Pallavi A R
Assistant professor
Dept. of CSE-DS
ATMECE, Mysuru

File-System Interface

Objectives

- To explain the function of file systems
- To describe the interfaces to file systems
- To discuss file-system design tradeoffs, including access methods, file sharing, file locking, and directory structures
- To explore file-system protection

File Concept

- Contiguous logical address space
- Types:
 - – Data
 - numeric
 - character
 - binary
 - – Program
 - Contents defined by file's creator
- – Many types
 - Consider text file, source file, executable file

File Attributes

- **Name** – only information kept in human-readable form
- **Identifier** – unique tag (number) identifies file within file system
- **Type** – needed for systems that support different types
- **Location** – pointer to file location on device
- **Size** – current file size
- **Protection** – controls who can do reading, writing, executing
- **Time, date, and user identification** – data for protection, security, and usage monitoring
- Information about files are kept in the directory structure, which is maintained on the disk
- Many variations, including extended file attributes such as file checksum
- Information kept in the directory structure

File info Window on Mac OS X



File Operations

- File is an abstract data type
- Create
- Write –at write pointer location
- Read –at read pointer location
- Reposition within file - seek
- Delete
- Truncate
- Open(Fi) – search the directory structure on disk for entry Fi, and move the content of entry to memory
- Close (Fi) – move the content of entry Fi in memory to directory structure on disk

Open Files

- Several pieces of data are needed to manage open files:
 - – Open-file table: tracks open files
 - – File pointer: pointer to last read/write location, per process that has the file open
 - – File-open count: counter of number of times a file is open – to allow
 - removal of data from open-file table when last processes closes it
 - – Disk location of the file: cache of data access information
 - – Access rights: per-process access mode information

Open File Locking

- Provided by some operating systems and file systems
 - – Similar to reader-writer locks
 - – Shared lock similar to reader lock – several processes can acquire concurrently
 - – Exclusive lock similar to writer lock
- Mediates access to a file
- Mandatory or advisory:
 - – Mandatory – access is denied depending on locks held and requested
 - – Advisory – processes can find status of locks and decide what to do

File Locking Example – Java API

```
import java.io.*;
import java.nio.channels.*;
public class LockingExample {
    public static final boolean EXCLUSIVE = false;
    public static final boolean SHARED = true;
    public static void main(String arsg[]) throws IOException {
        FileLock sharedLock = null;
        FileLock exclusiveLock = null;
        try {
            RandomAccessFile raf = new RandomAccessFile("file.txt", "rw");
            // get the channel for the file
            FileChannel ch = raf.getChannel();
            // this locks the first half of the file - exclusive
            exclusiveLock = ch.lock(0, raf.length()/2, EXCLUSIVE);
            /** Now modify the data . . . */
            // release the lock
            exclusiveLock.release();
        }
    }
}
```

TFile Locking Example – Java API (Cont.)

```
// this locks the second half of the file - shared
sharedLock = ch.lock(raf.length()/2+1, raf.length(),
                    SHARED);
/** Now read the data . . . */
// release the lock
sharedLock.release();
} catch (java.io.IOException ioe) {
    System.err.println(ioe);
}finally {
    if (exclusiveLock != null)
        exclusiveLock.release();
    if (sharedLock != null)
        sharedLock.release();
}
}
```

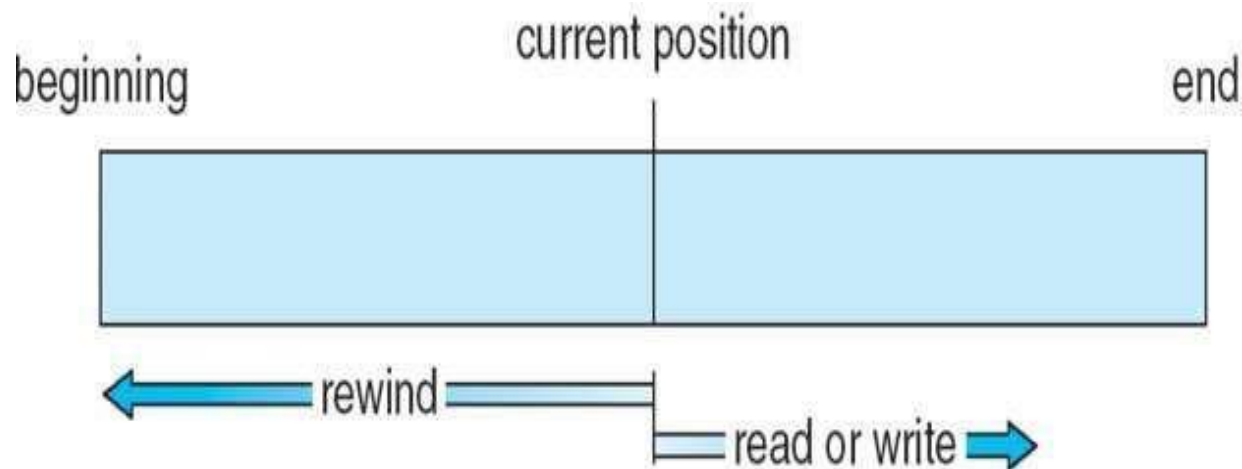
File Types – Name, Extension

file type	usual extension	function
executable	exe, com, bin or none	ready-to-run machine-language program
object	obj, o	compiled, machine language, not linked
source code	c, cc, java, pas, asm, a	source code in various languages
batch	bat, sh	commands to the command interpreter
text	txt, doc	textual data, documents
word processor	wp, tex, rtf, doc	various word-processor formats
library	lib, a, so, dll	libraries of routines for programmers
print or view	ps, pdf, jpg	ASCII or binary file in a format for printing or viewing
archive	arc, zip, tar	related files grouped into one file, sometimes compressed, for archiving or storage
multimedia	mpeg, mov, rm, mp3, avi	binary file containing audio or A/V information

File Structure

- None - sequence of words, bytes
- Simple record structure
 - – Lines
 - – Fixed length
 - – Variable length
- Complex Structures
 - – Formatted document
 - – Relocatable load file
- Can simulate last two with first method by inserting appropriate control characters
- Who decides:
 - – Operating system
 - – Program

Sequential-access File



Access Methods

- **Sequential Access**

```
read next
write next
reset
no read after last write
      (rewrite)
```

- **Direct Access** — file is fixed length **logical records**

```
read n
write n
position to n
      read next
      write next
rewrite n
```

n = **relative block number**

- Relative block numbers allow OS to decide where file should be placed
 - See **allocation problem** in Ch 12

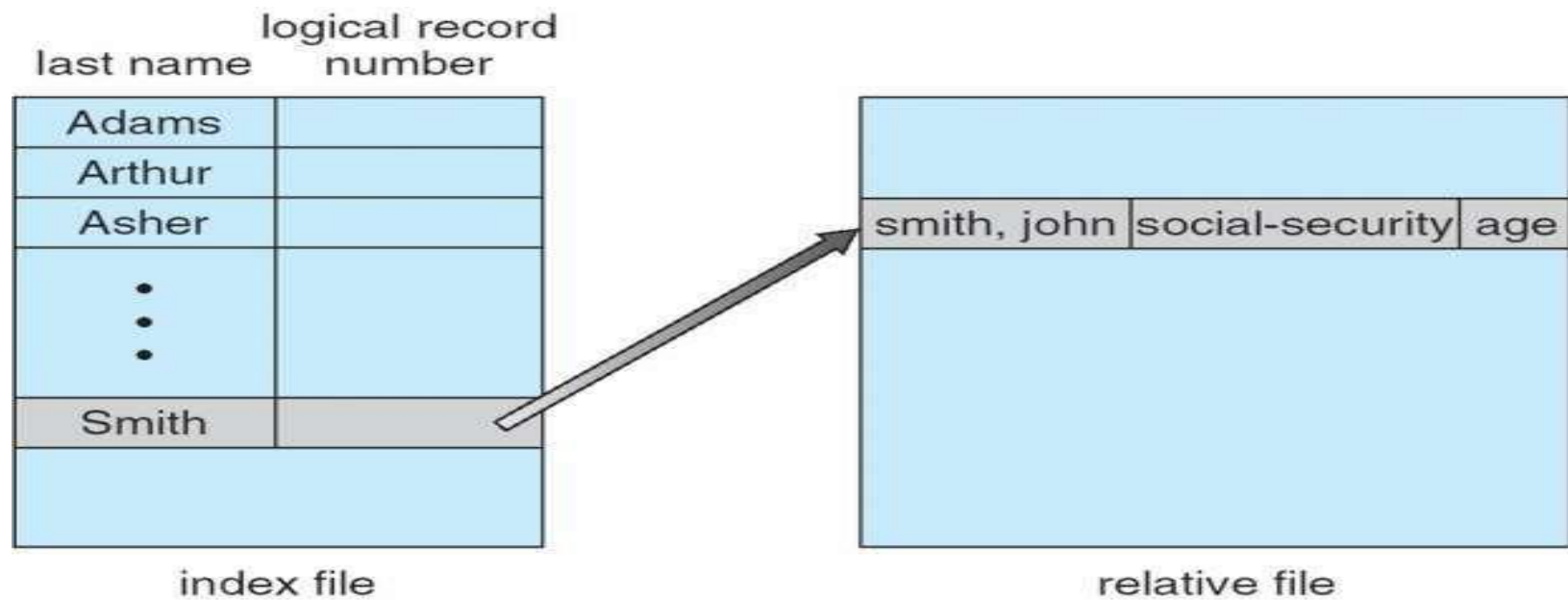
Simulation of Sequential Access on Direct-access File

sequential access	implementation for direct access
<i>reset</i>	<i>cp = 0;</i>
<i>read next</i>	<i>read cp;</i> <i>cp = cp + 1;</i>
<i>write next</i>	<i>write cp;</i> <i>cp = cp + 1;</i>

Other Access Methods

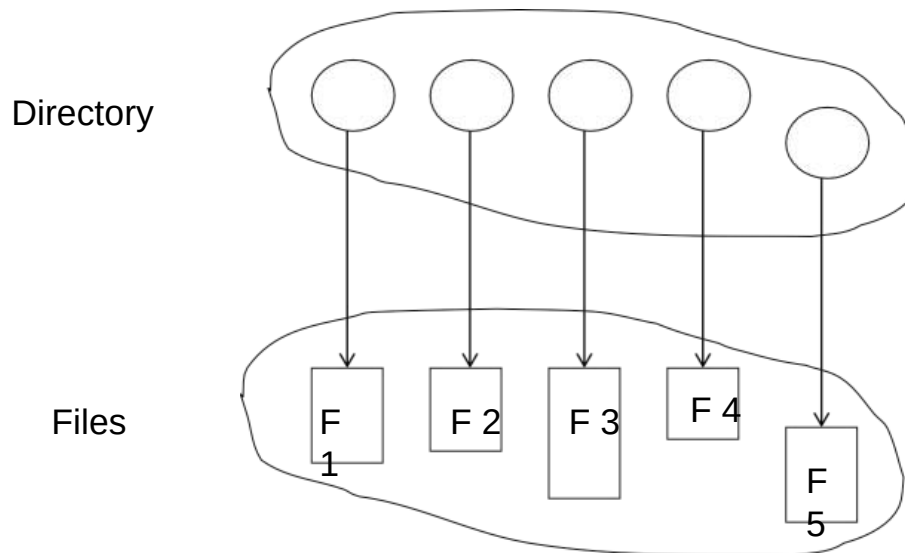
- Can be built on top of base methods
- General involve creation of an index for the file
- Keep index in memory for fast determination of location of data to be
- operated on (consider UPC code plus record of data about that item)
- If too large, index (in memory) of the index (on disk)
- IBM indexed sequential-access method (ISAM)
 - – Small master index, points to disk blocks of secondary index
 - – File kept sorted on a defined key
 - – All done by the OS
- • VMS operating system provides index and relative files as another example

Example of Index and Relative Files



Directory Structure

- A collection of nodes containing information about all files

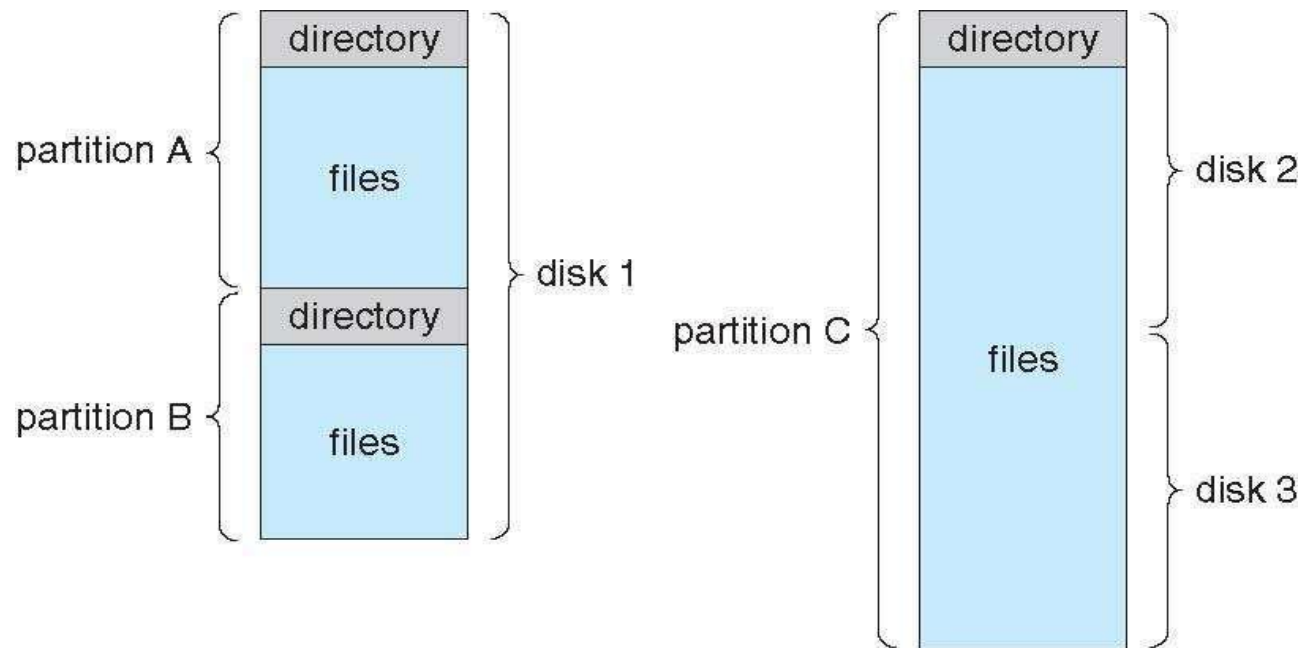


Both the directory structure and the files reside on disk

Disk Structure

- Disk can be subdivided into partitions
- Disks or partitions can be RAID protected against failure
- Disk or partition can be used raw – without a file system, or formatted with a file system
- Partitions also known as minidisks, slices
- Entity containing file system known as a volume
- Each volume containing file system also tracks that file system's info in device directory or volume table of contents
- As well as general-purpose file systems there are many special-purpose file systems, frequently all within the same operating system or computer

A Typical File-system Organization



Types of File Systems

- We mostly talk of general-purpose file systems
- But systems frequently have many file systems, some general- and some special-purpose
- Consider Solaris has
 - – tmpfs – memory-based volatile FS for fast, temporary I/O
 - – objfs – interface into kernel memory to get kernel symbols for debugging
 - – ctfs – contract file system for managing daemons
 - – lofs – loopback file system allows one FS to be accessed in place of another
 - – procfs – kernel interface to process structures
 - – ufs, zfs – general purpose file systems

Operations Performed on Directory

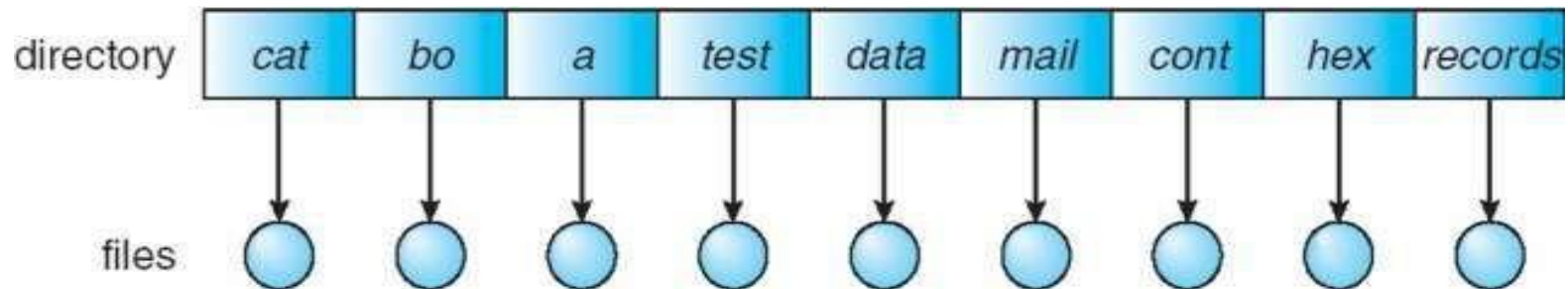
- Search for a file
- Create a file
- Delete a file
- List a directory
- Rename a file
- Traverse the file system

Directory Organization

- The directory is organized logically to obtain
- Efficiency – locating a file quickly
- Naming – convenient to users
 - – Two users can have same name for different files
 - – The same file can have several different names
- Grouping – logical grouping of files by
- properties, (e.g., all Java programs, all games,...)

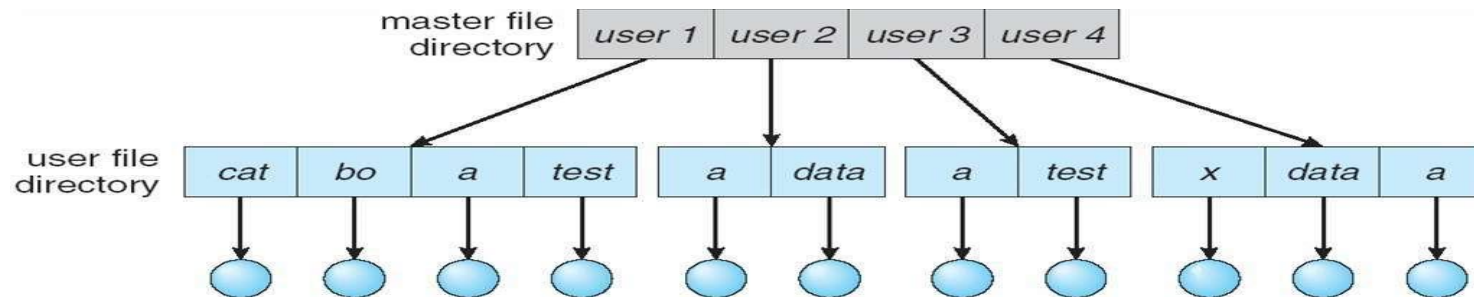
Single-Level Directory

- A single directory for all users
- Naming problem
- Grouping problem



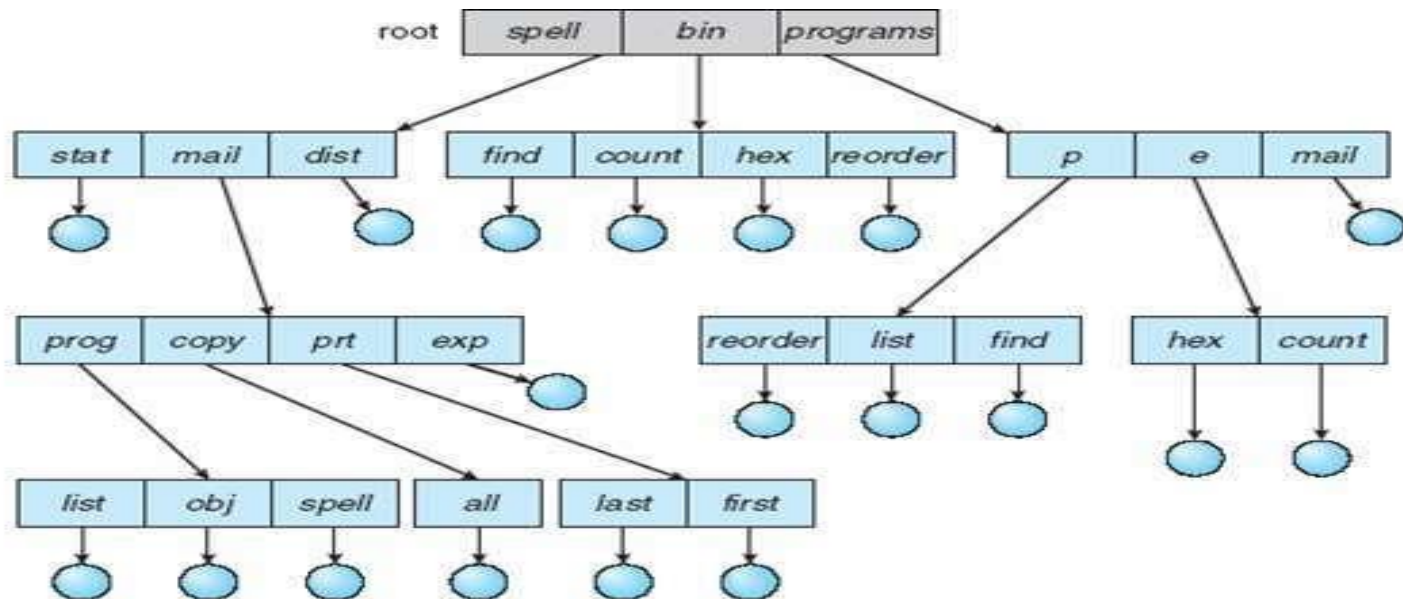
Two-Level Directory

- Separate directory for each user



- n Path name
- n Can have the same file name for different user
- n Efficient searching
- n No grouping capability

Tree-Structured Directories



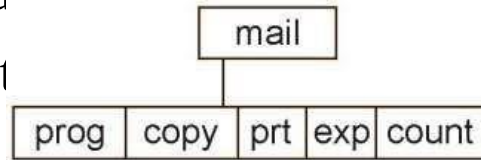
Need For Page Replacement

Tree-Structured Directories (Cont.)

- Efficient searching
- Grouping Capability
- Current directory (working directory)
 - – *cd /spell/mail/prog*
 - – *type list*

Tree-Structured Directories (Cont)

- Absolute or relative path name
- Creating a new file is done in current directory
- Delete a file
- `rm <file-name>`
- Creating a new subdirectory is done in current directory
- `mkdir <dir-name>`
- Example: if in current directory

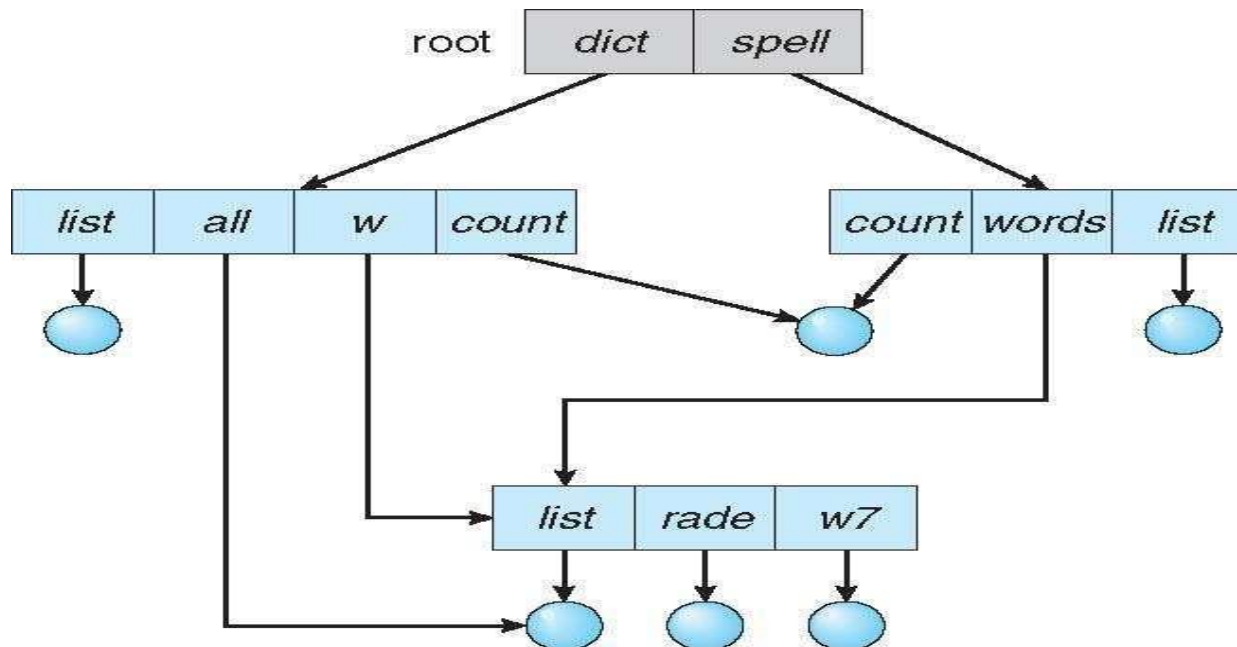


```

graph TD
    mail[mail] --- row
    subgraph row [ ]
        direction LR
        prog[prog] --- copy[copy] --- prt[prt] --- exp[exp] --- count[count]
    end
  
```
- ☐ Deleting “mail” ☒ deleting the entire subtree rooted by “mail”

Acyclic-Graph Directories

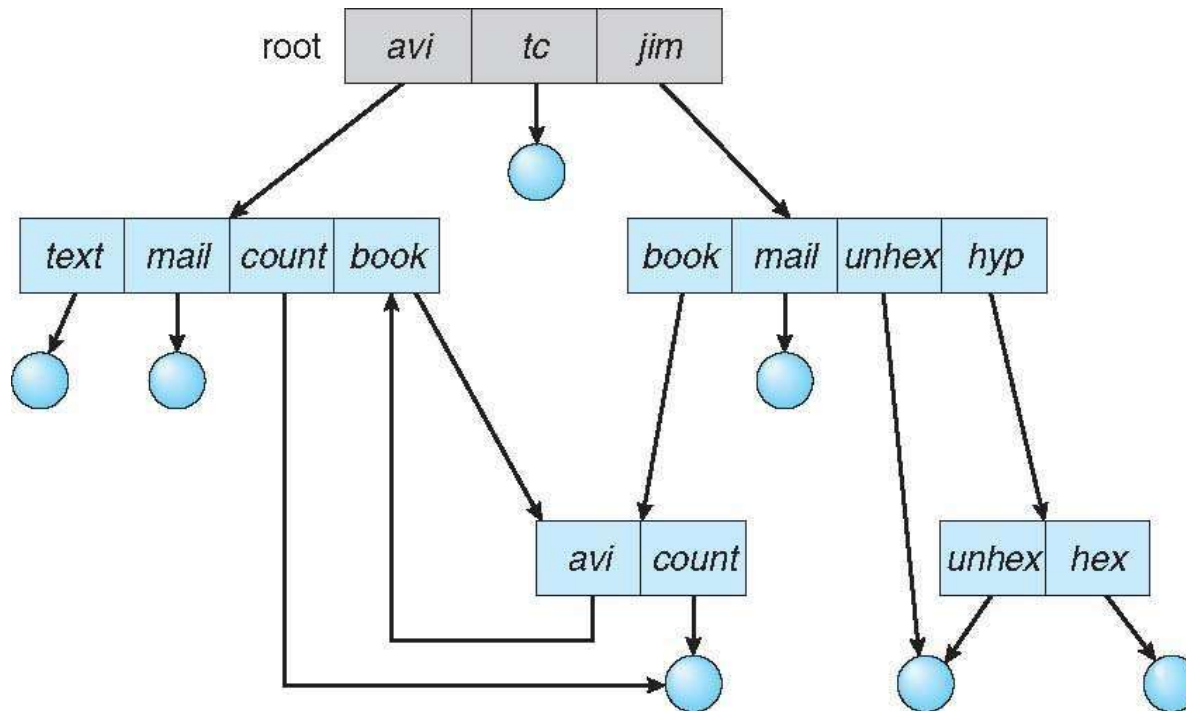
- Have shared subdirectories and files



Acyclic-Graph Directories (Cont.)

- Two different names (aliasing)
- If dict deletes list   dangling pointer
- Solutions:
 - – Backpointers, so we can delete all pointers
 - Variable size records a problem
 - – Backpointers using a daisy chain organization
 - – Entry-hold-count solution
 - New directory entry type
 - – Link – another name (pointer) to an existing file
 - – Resolve the link – follow pointer to locate the file

General Graph Directory

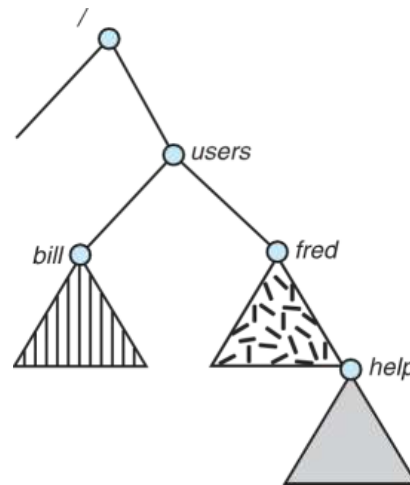


General Graph Directory (Cont.)

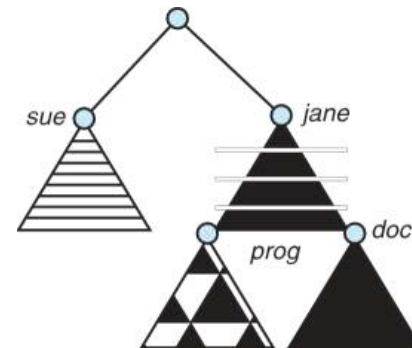
- How do we guarantee no cycles?
 - – Allow only links to file not subdirectories
 - – Garbage collection
 - – Every time a new link is added use a cycle detection
- algorithm to determine whether it is OK

File System Mounting

- A file system must be mounted before it can be accessed
- A unmounted file system (i.e., Fig. 11-11(b)) is mounted at a mount point

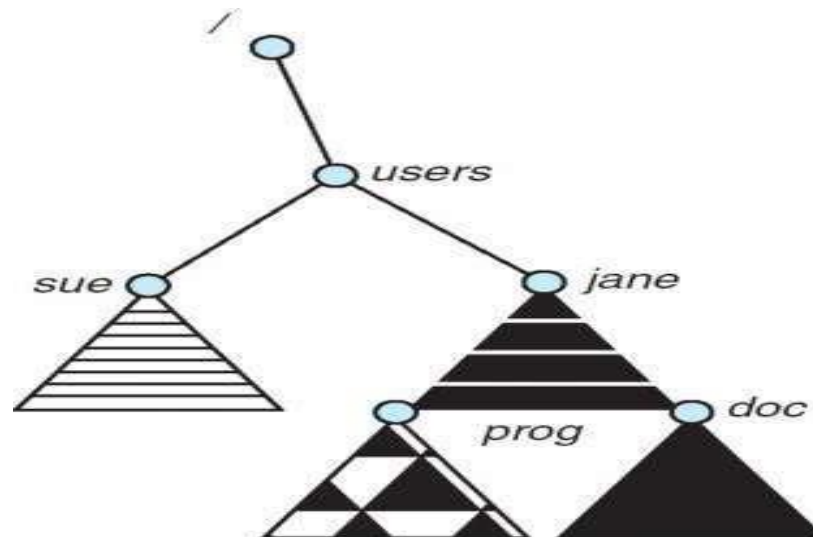


(a)



(b)

Mount Point



File Sharing

- Sharing of files on multi-user systems is desirable
- Sharing may be done through a protection scheme
- On distributed systems, files may be shared across a network
- Network File System (NFS) is a common distributed file-sharing method
- If multi-user system
 - – User IDs identify users, allowing permissions and protections to be per-user
 - – Group IDs allow users to be in groups, permitting group access rights
- – Owner of a file / directory
- – Group of a file / directory

File Sharing – Remote File Systems

- Uses networking to allow file system access between systems
 - – Manually via programs like FTP
 - – Automatically, seamlessly using distributed file systems
 - – Semi automatically via the world wide web
- • Client-server model allows clients to mount remote file systems from servers
 - – Server can serve multiple clients
 - – Client and user-on-client identification is insecure or complicated
 - – N F S is standard UNIX client-server file sharing protocol
 - – C I F S is standard Windows protocol
 - – Standard operating system file calls are translated into remote calls
- • Distributed Information Systems (distributed naming services) such as LDAP, DNS,
- NIS, Active Directory implement unified access to information needed for remote computing

File Sharing – Failure Modes

- All file systems have failure modes
 - – For example corruption of directory structures or other non-user data, called metadata
- Remote file systems add new failure modes, due to network failure, server failure
- Recovery from failure can involve state information about status of each remote request
- Stateless protocols such as NFS v3 include all information in each request, allowing easy recovery but less security

File Sharing – Consistency Semantics

- Specify how multiple users are to access a shared file simultaneously
- – Similar to Ch 5 process synchronization algorithms
- Tend to be less complex due to disk I/O and network latency (for remote file systems)
- – Andrew File System (AFS) implemented complex remote file sharing semantics
- – Unix file system (UFS) implements:
- Writes to an open file visible immediately to other users of the same open file
- Sharing file pointer to allow multiple users to read and write concurrently
- – AFS has session semantics
- Writes only visible to sessions starting after the file is closed

Protection

- File owner/creator should be able to control:
 - what can be done
 - by whom
- Types of access
 - Read
 - Write
 - Execute
 - Append
 - Delete
 - List

Access Lists and Groups

- Mode of access: read, write, execute
- Three classes of users on Unix / Linux

a) owner access	7	⇒	RWX 1 1 1
b) group access	6	⇒	RWX 1 1 0
c) public access	1	⇒	RWX 0 0 1

- Ask manager to create a group (unique name), say G, and add some users to the group.
- For a particular file (say *game*) or subdirectory, define an appropriate access.



Attach a group to a file

chgrp G game

A Sample UNIX Directory Listing

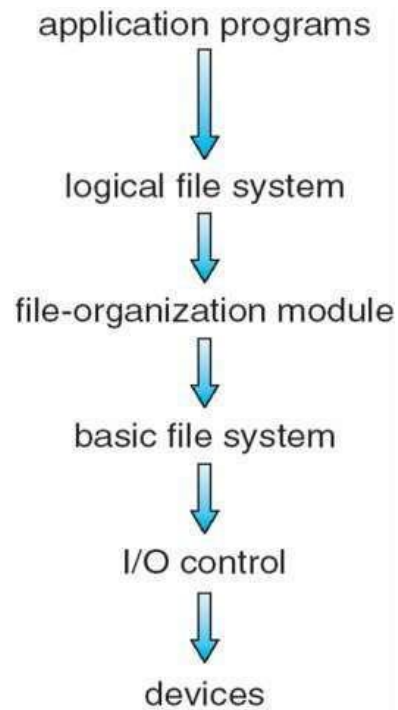
-rw-rw-r--	1	pbg	staff	31200	Sep 3 08:30	intro.ps
drwx-----	5	pbg	staff	512	Jul 8 09:33	private/
drwxrwxr-x	2	pbg	staff	512	Jul 8 09:35	doc/
drwxrwx---	2	pbg	student	512	Aug 3 14:13	student-proj/
-rw-r--r--	1	pbg	staff	9423	Feb 24 2003	program.c
-rwxr-xr-x	1	pbg	staff	20471	Feb 24 2003	program
drwx--x--x	4	pbg	faculty	512	Jul 31 10:31	lib/
drwx-----	3	pbg	staff	1024	Aug 29 06:52	mail/
drwxrwxrwx	3	pbg	staff	512	Jul 8 09:35	test/

File System Implementation

File-System Structure

- File structure
 - – Logical storage unit
 - – Collection of related information
- File system resides on secondary storage (disks)
 - – Provided user interface to storage, mapping logical to physical
 - – Provides efficient and convenient access to disk by allowing data to be stored, located retrieved easily
- Disk provides in-place rewrite and random access
 - – I/O transfers performed in blocks of sectors (usually 512 bytes)
- File control block – storage structure consisting of information about a file

Layered File System



File System Layers

- Device drivers manage I/O devices at the I/O control layer
- Given commands like “read drive1, cylinder 72, track 2,sector 10, into memory location 1060” outputs low-level hardware specific commands to hardware controller
- Basic file system given command like “retrieve block 123 ” translates to device driver
- Also manages memory buffers and caches (allocation, freeing, replacement)
- Buffers hold data in transit

File System Layers (Cont.)

- Logical file system manages metadata information
 - – Translates file name into file number, file handle, location by maintaining file control blocks (inodes in UNIX)
 - – Directory management
 - – Protection
- Layering useful for reducing complexity and redundancy, but adds overhead and can decrease performance
 - Translates file name into file number, file handle, location by maintaining file control

File System Layers (Cont.)

- Many file systems, sometimes many within an operating system
- Each with its own format (CD-ROM is ISO 9660; Unix has UFS, FFS; Windows has FAT, FAT32, NTFS as well as floppy, CD, DVD Blu-ray, Linux has more than 40 types, with extended file system ext2 and ext3 leading; plus distributed file systems, etc.)
- New ones still arriving – ZFS, GoogleFS, Oracle ASM,FUSE

File-System Implementation

- We have system calls at the API level, but how do we implement their functions?
- – On-disk and in-memory structures
- **Boot control block** contains info needed by system to boot OS from that volume
- – Needed if volume contains OS, usually first block of volume
- **Volume control block** (superblock, master file table) contains volume details
- – Total # of blocks, # of free blocks, block size, free block

File-System Implementation (Cont.)

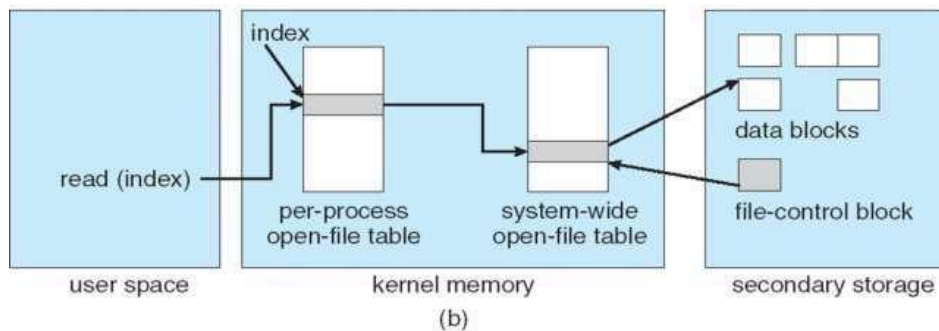
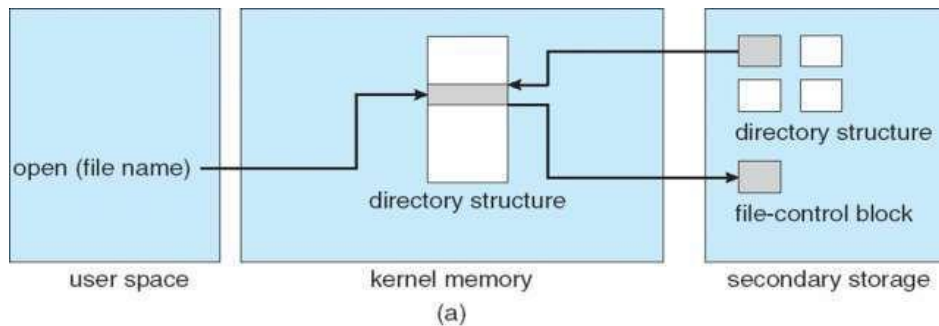
- Per-file File Control Block (FCB) contains many details about the file
 - – inode number, permissions, size, dates
 - – NFTS stores into in master file table using relational DB structure

file permissions
file dates (create, access, write)
file owner, group, ACL
file size
file data blocks or pointers to file data blocks

In-Memory File System Structures

- Mount table storing file system mounts, mount
- points, file system types
- The following figure illustrates the necessary file
- system structures provided by the operating systems
- Figure 12-3(a) refers to opening a file
- Figure 12-3(b) refers to reading a file
- Plus buffers hold data blocks from secondary storage

In-Memory File System Structures



Partitions and Mounting

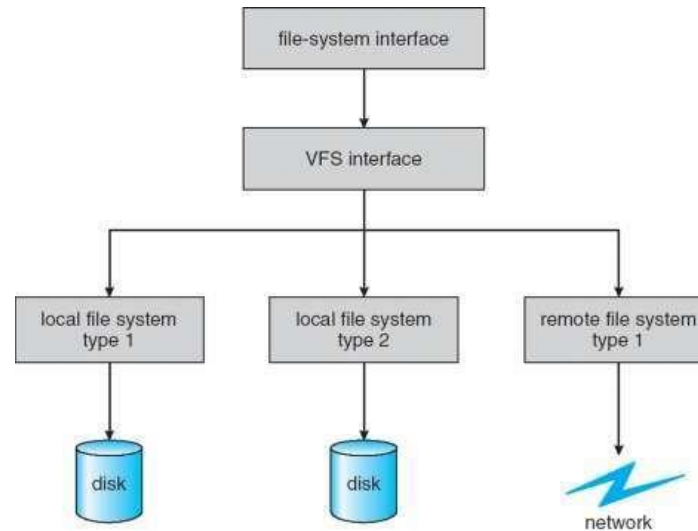
- Partition can be a volume containing a file system (“cooked”) or raw
 - – just a sequence of blocks with no file system
- Boot block can point to boot volume or boot loader set of blocks that contain enough code to know how to load the kernel from the file system
 - – Or a boot management program for multi-os booting
- Root partition contains the OS, other partitions can hold other Oses, other file systems, or be raw
 - – Mounted at boot time

Virtual File Systems

- Virtual File Systems (VFS) on Unix provide an object-oriented way of implementing file systems
- VFS allows the same system call interface (the API) to be used for different types of file systems
 - Separates file-system generic operations from implementation details
 - Implementation can be one of many file systems types, or network file system
- Implements vnodes which hold inodes or network file details
 - Then dispatches operation to appropriate file system

Virtual File Systems (Cont.)

- The API is to the VFS interface, rather than any specific type of file system



Virtual File System Implementation

- For example, Linux has four object types:
 - – inode, file, superblock, dentry
- VFS defines set of operations on the objects that must be implemented
 - – Every object has a pointer to a function table
- Function table has addresses of routines to implement that function on that object
 - • For example:
 - int open(. . .)— Open a file
 - int close(. . .)— Close an already-open file
 - ssize_t read(. . .)— Read from a file

Directory Implementation

- Linear list of file names with pointer to the data blocks
 - – Simple to program
 - – Time-consuming to execute
- • Linear search time
- • Could keep ordered alphabetically via linked list or use B+ tree
- • Hash Table – linear list with hash data structure
 - – Decreases directory search time
 - – Collisions – situations where two file names hash to the same location

Allocation Methods - Contiguous

- An allocation method refers to how disk blocks are allocated for files:
- Contiguous allocation – each file occupies set of contiguous blocks
 - – Best performance in most cases
 - – Simple – only starting location (block #) and length (number of blocks) are required
 - – Problems include finding space for file, knowing file size,
- external fragmentation, need for compaction off-line (downtime) or on-line

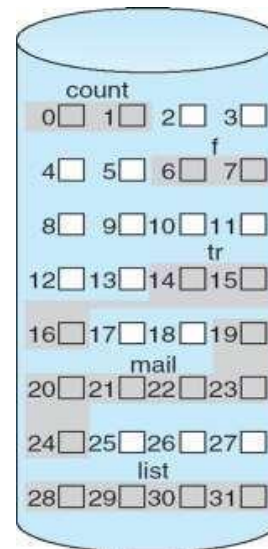
Contiguous Allocation

- Mapping from logical to
- physical

Q

LA/512

R



directory

file	start	length
count	0	2
tr	14	3
mail	19	6
list	28	4
f	6	2

- Block to be accessed = $Q +$
- starting address
- Displacement into block = R

Extent-Based Systems

- Many newer file systems (i.e., Veritas File System) use a modified contiguous allocation scheme
- Extent-based file systems allocate disk blocks in extents
- An extent is a contiguous block of disks
 - – Extents are allocated for file allocation
 - – A file consists of one or more extents

Allocation Methods - Linked

- Linked allocation – each file a linked list of blocks
 - – File ends at nil pointer
 - – No external fragmentation
 - – Each block contains pointer to next block
 - – No compaction, external fragmentation
 - – Free space management system called when new block needed
 - – Improve efficiency by clustering blocks into groups but increases internal fragmentation
 - – Reliability can be a problem
 - – Locating a block can take many I/Os and disk seeks

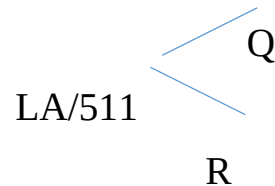
Allocation Methods – Linked (Cont.)

- FAT (File Allocation Table) variation
 - – Beginning of volume has table, indexed by block number
 - – Much like a linked list, but faster on disk and cacheable
 - – New block allocation simple

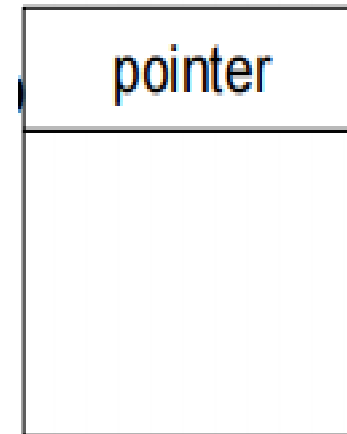
Linked Allocation

- Each file is a linked list of disk blocks: blocks may be
- not contiguous or scattered anywhere

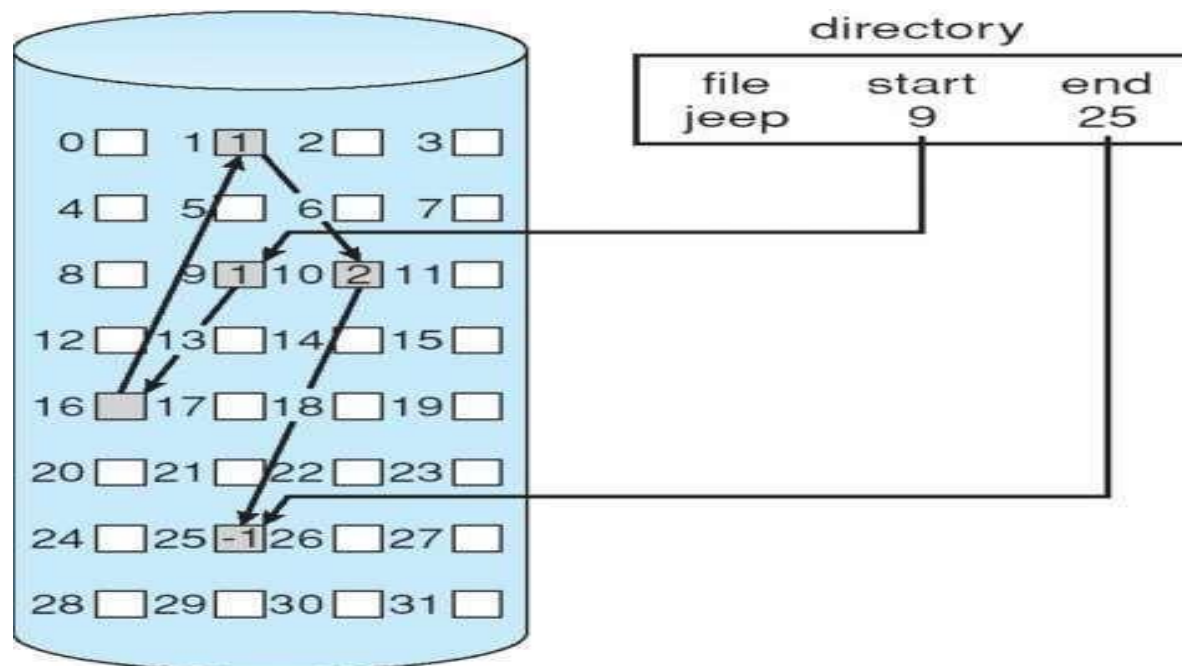
- Mapping



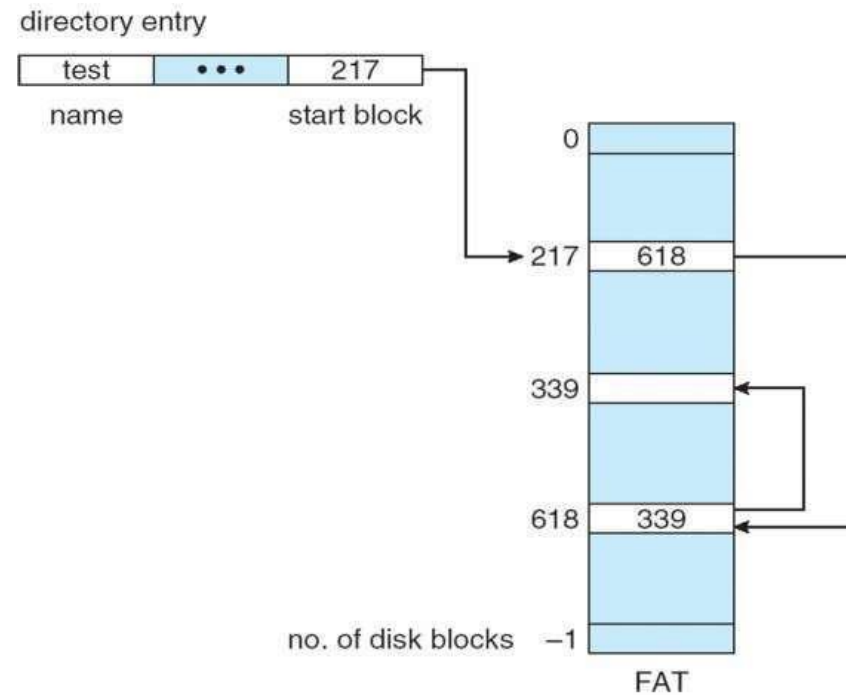
- Block to be accessed is the Qth block in the linked chain of blocks
- representing the file.
- Displacement into block = $R + 1$



Linked Allocation

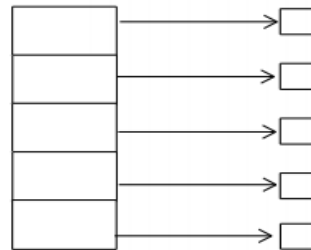


File-Allocation Table



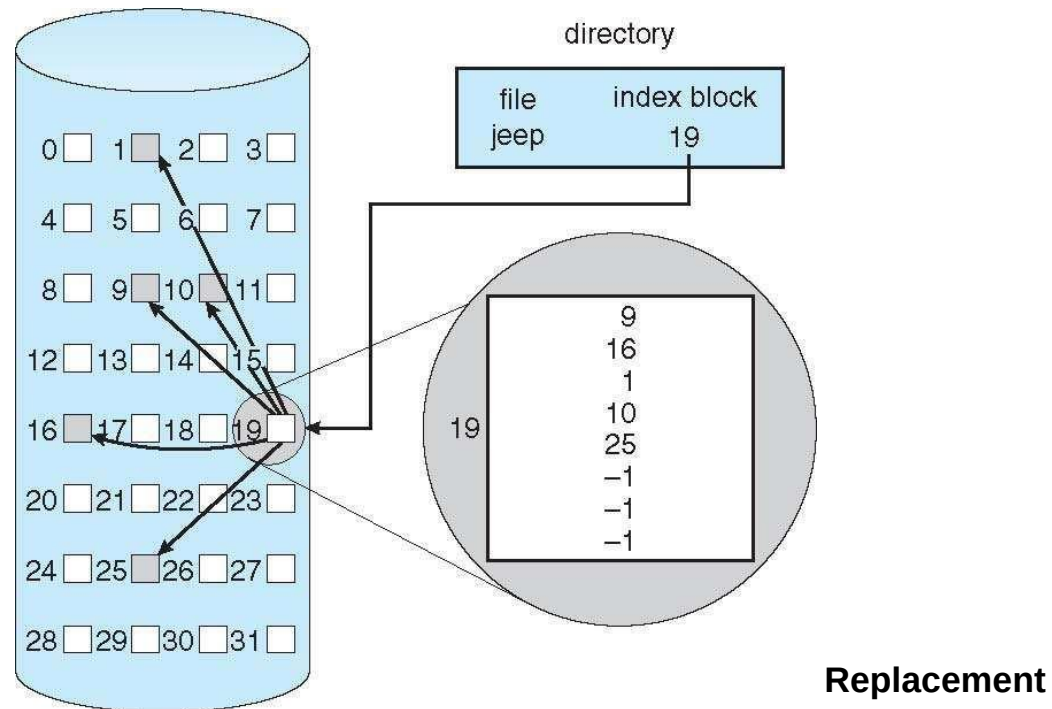
Allocation Methods -Indexed

- Indexed allocation
- – Each file has its own index bloc
- k(s) of pointers to its data blocks
- Logical view



index table

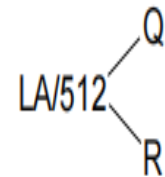
Example of Indexed Allocation



Indexed Allocation (Cont.)

- Need index table
- Random access
- Dynamic access without external fragmentation, but
- have overhead of index block
- Mapping from logical to physical in a file of
- maximum size of 256K bytes and block size of 512
- bytes. We need only 1 block for index table

Q = displacement into index table
R = displacement into block



Indexed Allocation – Mapping (Cont.)

- Mapping from logical to physical in a file of unbounded length (block size of 512 words)
- Linked scheme - Link blocks of index table (no limit on size)

$$LA / (512 \times 511) \begin{cases} Q_1 \\ R_1 \end{cases}$$

Q_1 = block of index table
 R_1 is used as follows:

$$R_1 / 512 \begin{cases} Q_2 \\ R_2 \end{cases}$$

Q_2 = displacement into block of index table
 R_2 displacement into block of file:

Indexed Allocation – Mapping (Cont.)

Two-level index (4K blocks could store 1,024 four-byte pointers in outer index $\rightarrow$ 1,048,567 data blocks and file size of up to 4GB)

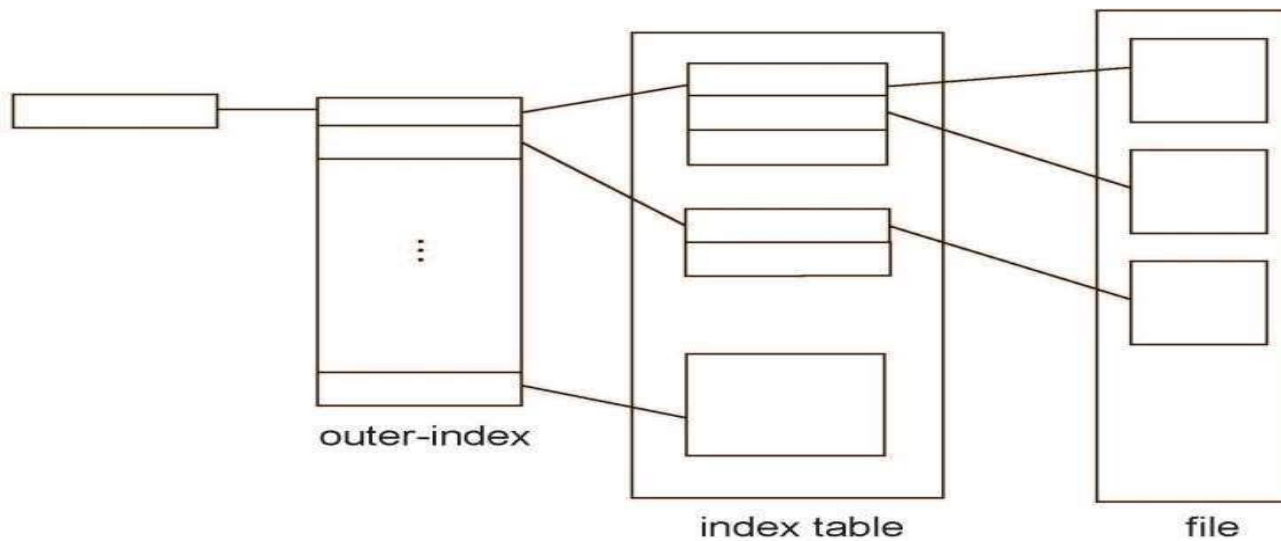
$$LA / (512 \times 512) \begin{matrix} \nearrow Q_1 \\ \searrow R_1 \end{matrix}$$

Q_1 = displacement into outer-index
 R_1 is used as follows:

$$R_1 / 512 \begin{matrix} \nearrow Q_2 \\ \searrow R_2 \end{matrix}$$

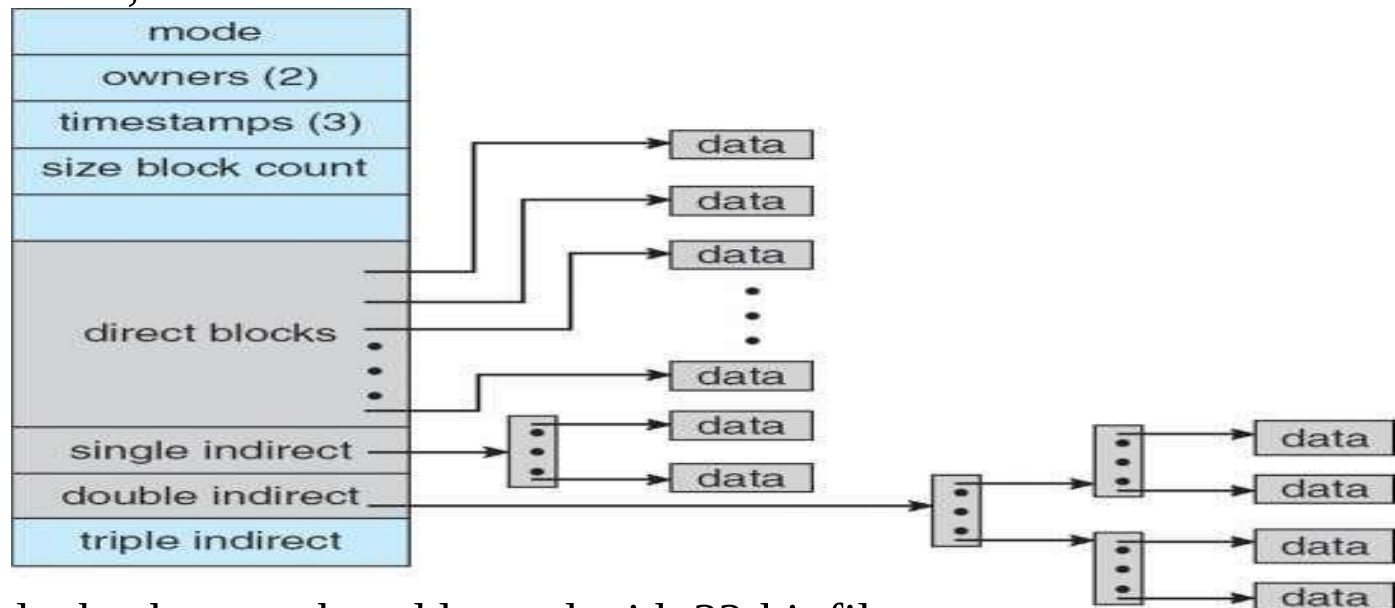
Q_2 = displacement into block of index table
 R_2 displacement into block of file:

Indexed Allocation – Mapping (Cont.)



Combined Scheme: UNIX UFS

- 4K bytes per block, 32-bit addresses



- More index blocks than can be addressed with 32-bit file p

Performance

- Best method depends on file access type
 - – Contiguous great for sequential and random
- Linked good for sequential, not random
- Declare access type at creation -> select either contiguous or linked
- Indexed more complex
 - – Single block access could require 2 index block reads then data block read
- – Clustering can help improve throughput, reduce CPU overhead

Performance (Cont.)

- Adding instructions to the execution path to save one
- disk I/O is reasonable
 - – Intel Core i7 Extreme Edition 990x (2011) at 3.46Ghz =
 - 159,000 MIPS
 - • http://en.wikipedia.org/wiki/Instructions_per_second
 - – Typical disk drive at 250 I/Os per second
 - • $159,000 \text{ MIPS} / 250 = 630$ million instructions during one disk I/O
 - – Fast SSD drives provide 60,000 IOPS
 - • $159,000 \text{ MIPS} / 60,000 = 2.65$ millions instructions during one disk I/O

Free-Space Management

- File system maintains free-space list to track available
- blocks/clusters



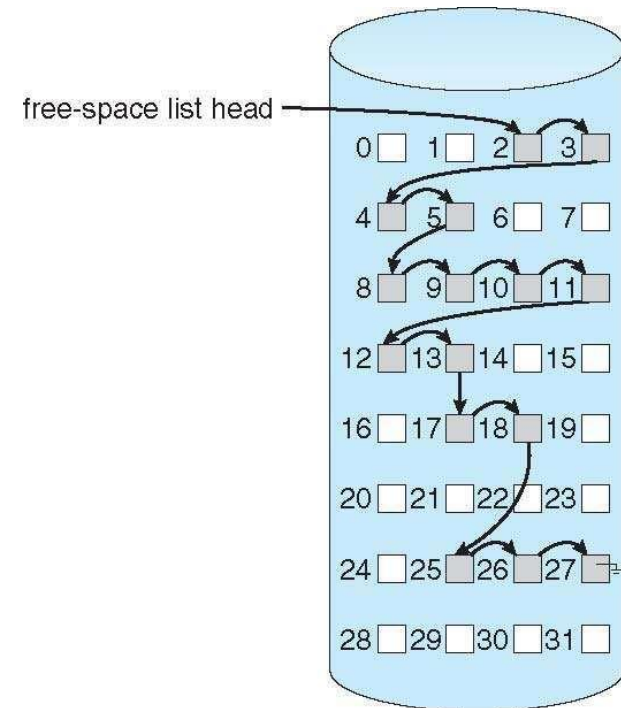
- Bit vector or bit map (n blocks)
- ... – (Using term “block” for simplicity) n-1
- $\text{bit}[i] = 1 \Rightarrow \text{block}[i] \text{ free}$
- $0 \Rightarrow \text{block}[i] \text{ occupied}$
- Block number calculation
- (number of bits per word) *
- (number of 0-value words) +
- offset of first 1 bit
- CPUs have instructions to return offset within word of first “1” bit

Free-Space Management (Cont.)

- Bit map requires extra space
- – Example:
- block size = 4KB = 212bytes
- disk size = 240 bytes (1 terabyte)
- $n = 240/212 = 228$ bits (or 32MB)
- if clusters of 4 blocks -> 8MB of memory
- Easy to get contiguous files

Linked Free Space List on Disk

- Linked list (free list)
- Cannot get contiguous space easily
- ☐ No waste of space
- ☐ No need to traverse the
- entire list (if # free blocks recorded)



Free-Space Management (Cont.)

- Grouping
 - – Modify linked list to store address of next $n-1$ free blocks in
 - first free block, plus a pointer to next block that contains
 - free-block-pointers (like this one)
- Counting
 - – Because space is frequently contiguously used and freed,
 - with contiguous-allocation allocation, extents, or clustering
 - Keep address of first free block and count of following free blocks
 - Free space list then has entries containing addresses and counts

Free-Space Management (Cont.)

- Space Maps
 - – Used in Z F S
 - – Consider meta-data I/O on very large file systems
- Full data structures like bit maps couldn't fit in memory -> thousands of I/Os
 - – Divides device space into metaslab units and manages metaslabs
- Given volume can contain hundreds of metaslabs
 - – Each metaslab has associated space map
- Uses counting algorithm
 - – But records to log file rather than file system
- Log of all block activity, in time order, in counting format
 - – Metaslab activity -> load space map into memory in balanced-treestructure, indexed by offset
- Replay log into that structure
- Combine contiguous free blocks into single entry

Mass-Storage Systems

- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Stable-Storage Implementation

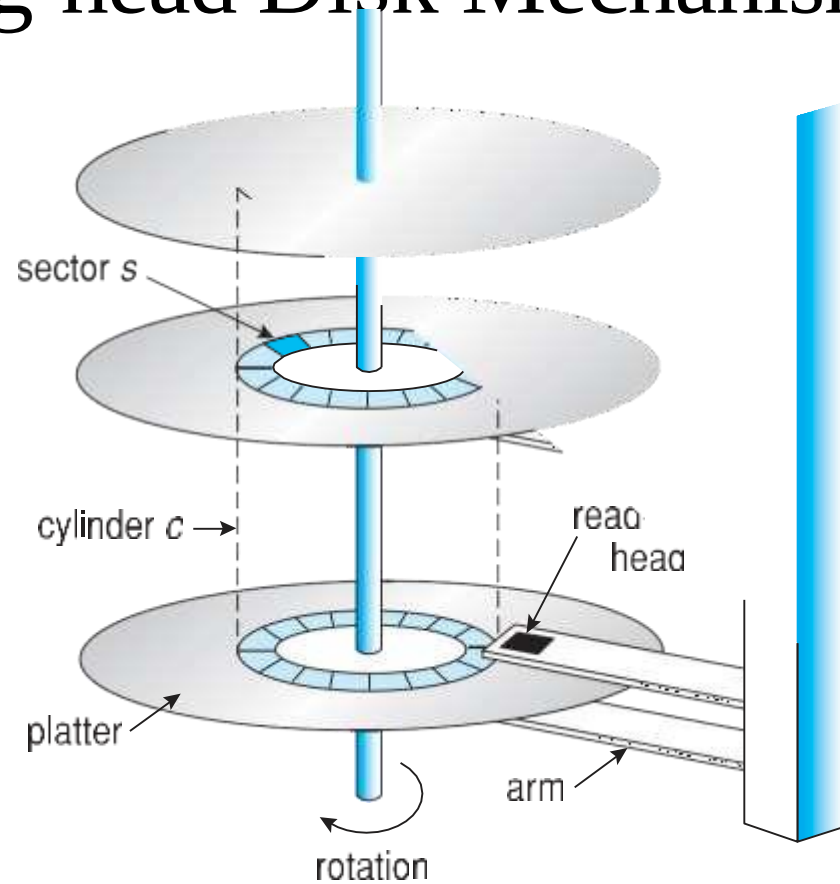
Objectives

- To describe the physical structure of secondary storage devices and its effects on the uses of the devices
- To explain the performance characteristics of mass-storage devices
- To evaluate disk scheduling algorithms
- To discuss operating-system services provided for mass storage, including RAID

Overview of Mass Storage Structure

- **Magnetic disks** provide bulk of secondary storage of modern computers
 - Drives rotate at 60 to 250 times per second
 - **Transfer rate** is rate at which data flow between drive and computer
 - **Positioning time** (**random-access time**) is time to move disk arm to desired cylinder (**seek time**) and time for desired sector to rotate under the disk head (**rotational latency**)
 - **Head crash** results from disk head making contact with the disk surface – That's bad
- Disks can be removable
- Drive attached to computer via **I/O bus**
 - Busses vary, including **EIDE**, **ATA**, **SATA**, **USB**, **Fibre Channel**, **SCSI**, **SAS**, **Firewire**
 - **Host controller** in computer uses bus to talk to **disk controller** built into drive or storage array

Moving-head Disk Mechanism



- Platters range from .85” to 14” (historically)
 - Commonly 3.5”, 2.5”, and 1.8”
- Range from 30GB to 3TB per drive
- Performance
 - Transfer Rate – theoretical – 6 Gb/sec
 - Effective Transfer Rate –real –1Gb/sec
 - Seek time from 3ms to 12ms – 9ms common for desktop drives
 - Average seek time measured or calculated based on 1/3 of tracks
 - Latency based on spindle speed
 - $1 / (\text{RPM} / 60) = 60 / \text{RPM}$
 - Average latency = $\frac{1}{2}$ latency

Spindle [rpm]	Average latency [ms]
4200	7.14
5400	5.56
7200	4.17
10000	3
15000	2

(From Wikipedia)

Hard Disk Performance

- **Access Latency** = **Average access time** = average seek time + average latency
 - For fastest disk $3\text{ms} + 2\text{ms} = 5\text{ms}$
 - For slow disk $9\text{ms} + 5.56\text{ms} = 14.56\text{ms}$
- Average I/O time = average access time + (amount to transfer / transfer rate) + controller overhead
- For example to transfer a 4KB block on a 7200 RPM disk with a 5ms average seek time, 1Gb/sec transfer rate with a .1ms controller overhead =
 - $5\text{ms} + 4.17\text{ms} + 0.1\text{ms} + \text{transfer time} =$
 - Transfer time = $4\text{KB} / 1\text{Gb/s} * 8\text{Gb} / \text{GB} * 1\text{GB} / 1024^2\text{KB} = 32 / (1024^2) = 0.031\text{ ms}$
 - Average I/O time for 4KB block = $9.27\text{ms} + .031\text{ms} = 9.301\text{ms}$

The First Commercial Disk Drive



1956
IBM RAMDAC computer
included the IBM Model 350
disk storage system

5M (7 bit) characters
50 x 24" platters
Access time = < 1 second

Solid-State Disks

- Nonvolatile memory used like a hard drive
 - – Many technology variations
- Can be more reliable than HDDs
- More expensive per MB
 - Maybe have shorter life span
- Less capacity
- But much faster
- Busses can be too slow -> connect directly to PCI for example
- No moving parts, so no seek time or rotational latency

Magnetic Tape

- Was early secondary-storage medium
 - Evolved from open spools to cartridges
- Relatively permanent and holds large quantities of data
- Access time slow
- Random access ~1000 times slower than disk
- Mainly used for backup, storage of infrequently-used data, transfer medium between systems
- Kept in spool and wound or rewound past read-write head
- Once data under head, transfer rates comparable to disk
 - 140MB/sec and greater
- 200GB to 1.5TB typical storage
- Common technologies are LTO- $\{3,4,5\}$ and T10000

Disk Structure

- Disk drives are addressed as large 1-dimensional arrays of **logical blocks**, where the logical block is the smallest unit of transfer
 - Low-level formatting creates **logical blocks** on physical media
- The 1-dimensional array of logical blocks is mapped into the sectors of the disk sequentially
 - Sector 0 is the first sector of the first track on the outermost cylinder
 - Mapping proceeds in order through that track, then the rest of the tracks in that cylinder, and then through the rest of the cylinders from outermost to innermost
 - Logical to physical address should be easy
 - Except for bad sectors
 - Non-constant # of sectors per track via constant angular velocity

Disk Attachment

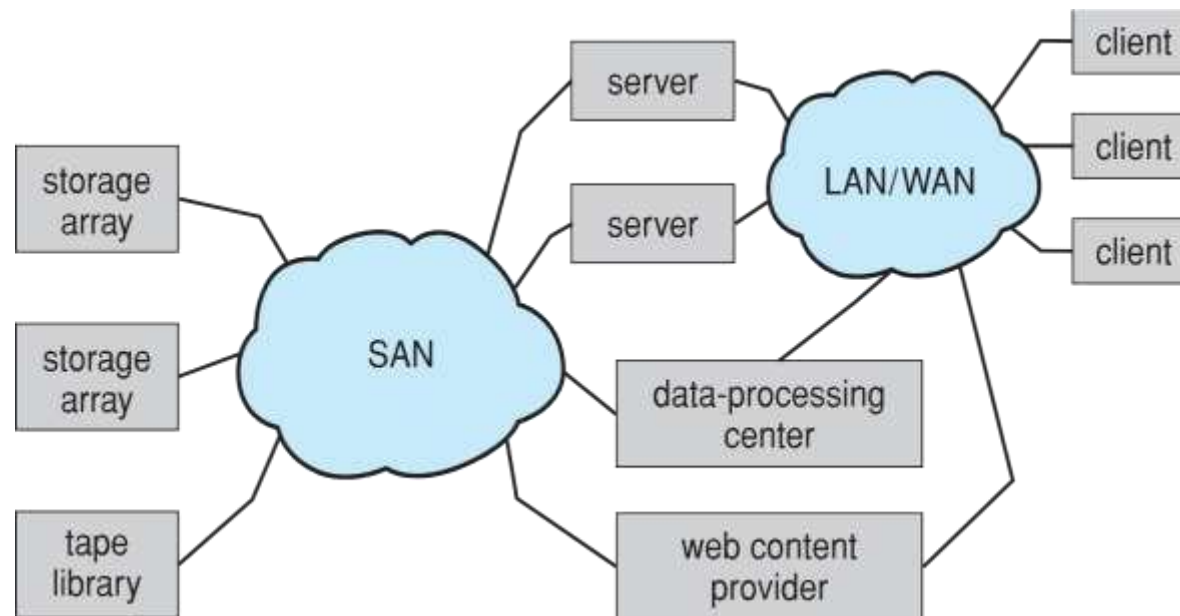
- Host-attached storage accessed through I/O ports talking to I/O busses
- SCSI itself is a bus, up to 16 devices on one cable, **SCSI initiator** requests operation and **SCSI targets** perform tasks
 - Each target can have up to 8 **logical units** (disks attached to device controller)
- FC is high-speed serial architecture
 - Can be switched fabric with 24-bit address space –the basis of **storage area networks (SANs)** in which many hosts attach to many storage units
- I/O directed to bus ID, device ID, logical unit (LUN)

Storage Array

- Can just attach disks, or arrays of disks
- Storage Array has controller(s), provides features to attached host(s)
 - Ports to connect hosts to array
 - Memory, controlling software (sometimes NVRAM, etc)
 - A few to thousands of disks
 - RAID, hot spares, hot swap (discussed later)
 - Shared storage -> more efficiency
 - Features found in some file systems
 - Snapshots, clones, thin provisioning, replication, deduplication, etc

Storage Area Network

- Common in large storage environments
- Multiple hosts attached to multiple storage arrays - flexible

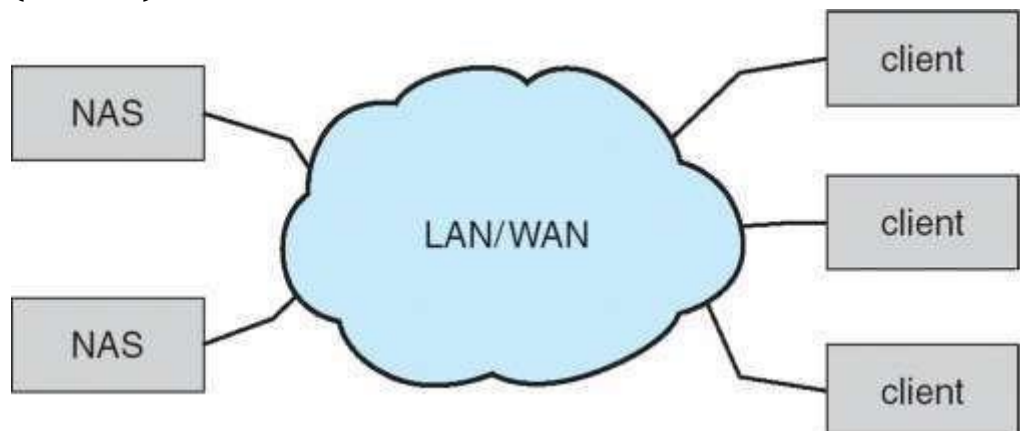


Storage Area Network (Cont.)

- SAN is one or more storage arrays
 - Connected to one or more Fibre Channel switches
- Hosts also attach to the switches
- Storage made available via **LUN Masking** from specific arrays to specific servers
- Easy to add or remove storage, add new host and allocate it storage
 - Over low-latency Fibre Channel fabric
- Why have separate storage networks and communications networks?
 - Consider iSCSI, FCOE

Network-Attached Storage

- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
 - Remotely attaching to file systems
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage over typically TCP or UDP on IP network
- **iSCSI** protocol uses IP network to carry the SCSI protocol
 - Remotely attaching to devices (blocks)



Disk Scheduling

- The operating system is responsible for using hardware efficiently — for the disk drives, this means having a fast access time and disk bandwidth
- Minimize seek time
- Seek time $\approx$ seek distance
- Disk **bandwidth** is the total number of bytes transferred, divided by the total time between the first request for service and the completion of the last transfer

- There are many sources of disk I/O request
 - OS
 - System processes
 - Users processes
- I/O request includes input or output mode, disk address, memory address, number of sectors to transfer
- OS maintains queue of requests, per disk or device
- Idle disk can immediately work on I/O request, busy disk means work must queue
 - Optimization algorithms only make sense when a queue exists

Disk Scheduling (Cont.)

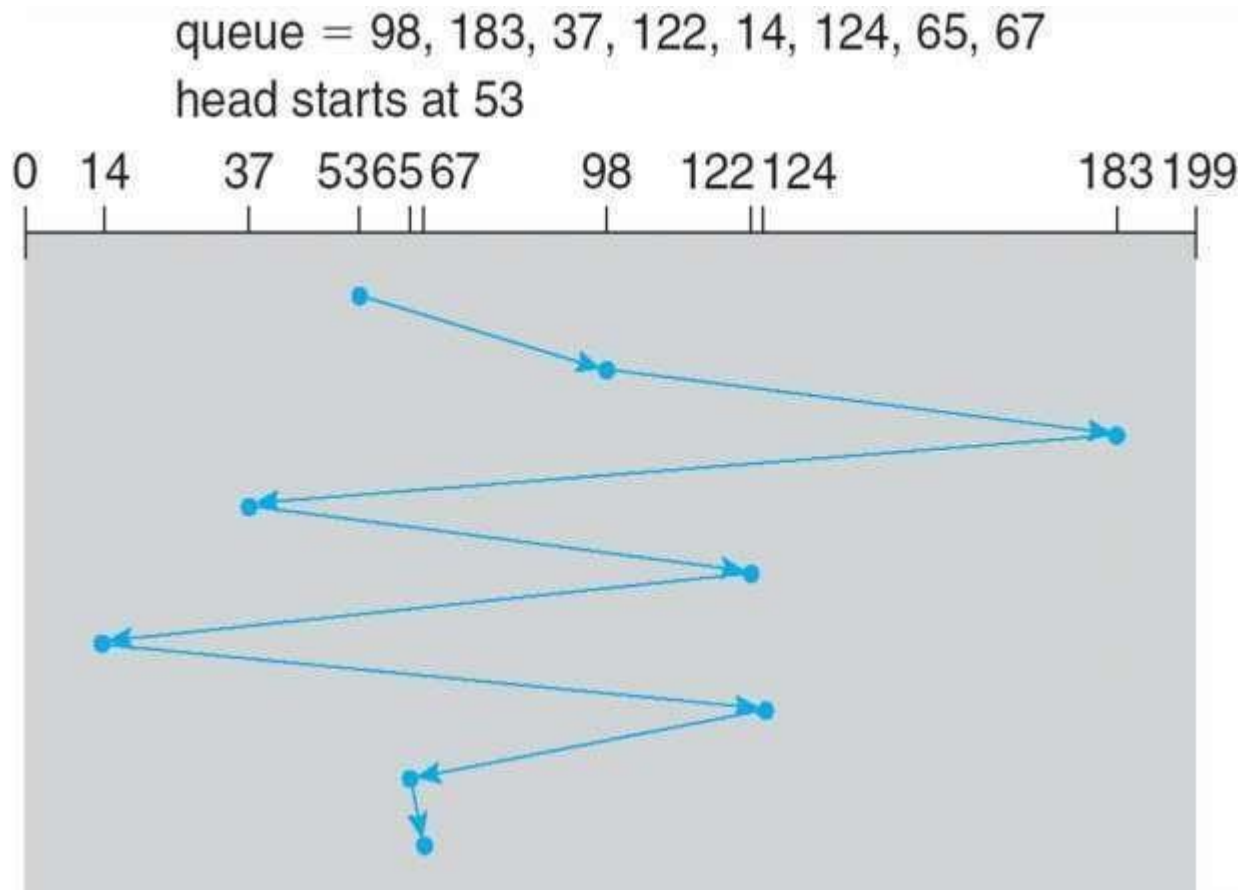
- Note that drive controllers have small buffers and can manage a queue of I/O requests (of varying “depth”)
- Several algorithms exist to schedule the servicing of disk I/O requests
- The analysis is true for one or many platters
- We illustrate scheduling algorithms with a request queue (0-199)

98, 183, 37, 122, 14, 124, 65, 67

Head pointer 53

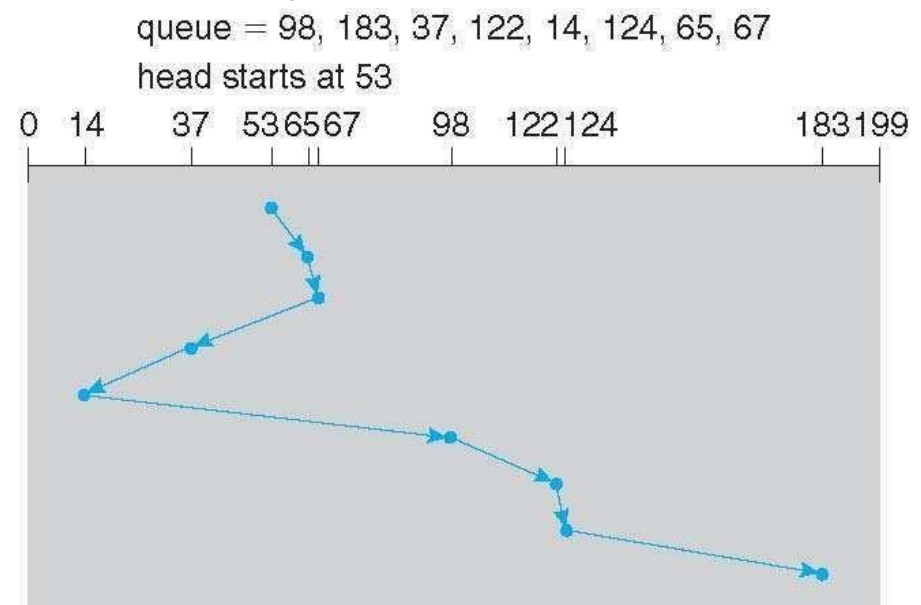
FCFS

Illustration shows total head movement of 640 cylinders



SSTF

- Shortest Seek Time First selects the request with the minimum seek time from the current head position
- SSTF scheduling is a form of SJF scheduling; may cause starvation of some requests
- Illustration shows total head movement of 236 cylinders



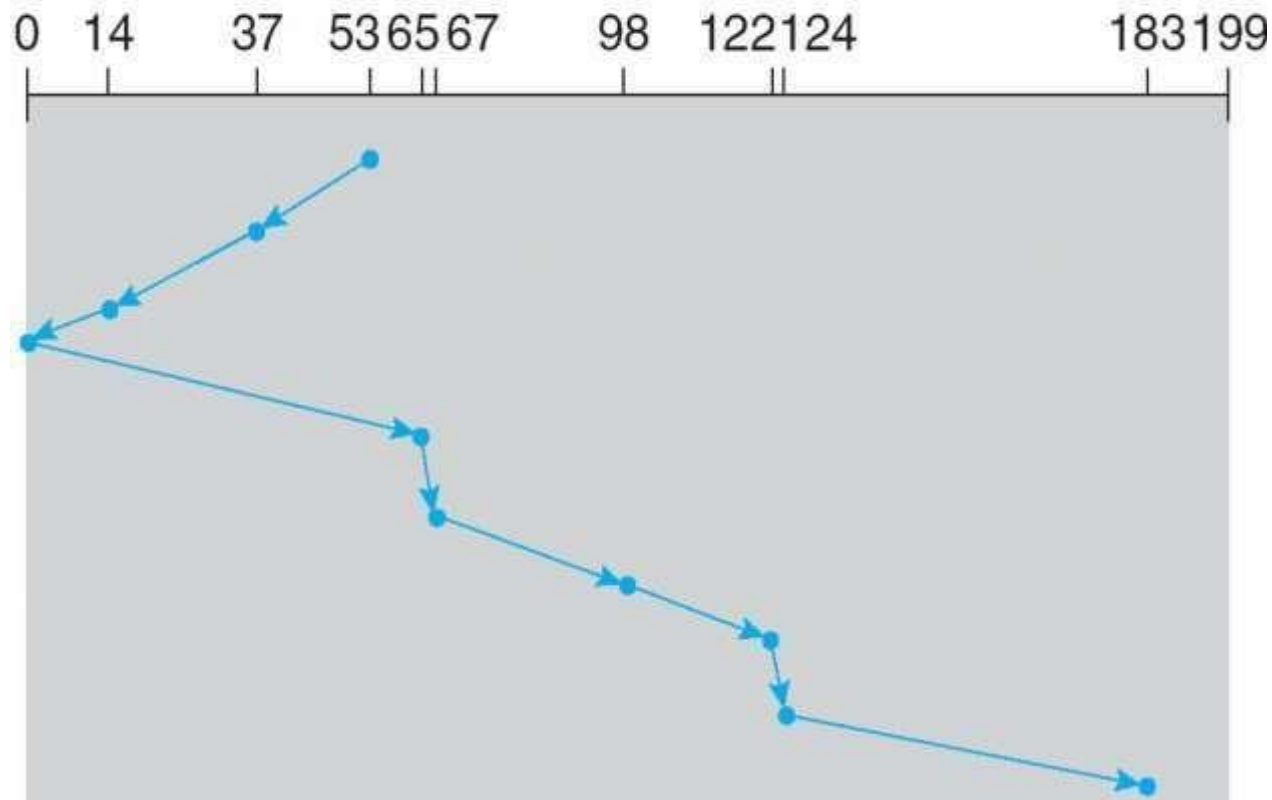
SCAN

- The disk arm starts at one end of the disk, and moves toward the other end, servicing requests until it gets to the other end of the disk, where the head movement is reversed and servicing continues.
- **SCAN algorithm** Sometimes called the **elevator algorithm**
- Illustration shows total head movement of 236 cylinders
- But note that if requests are uniformly dense, largest density at other end of disk and those wait the longest

SCAN (Cont.)

queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53

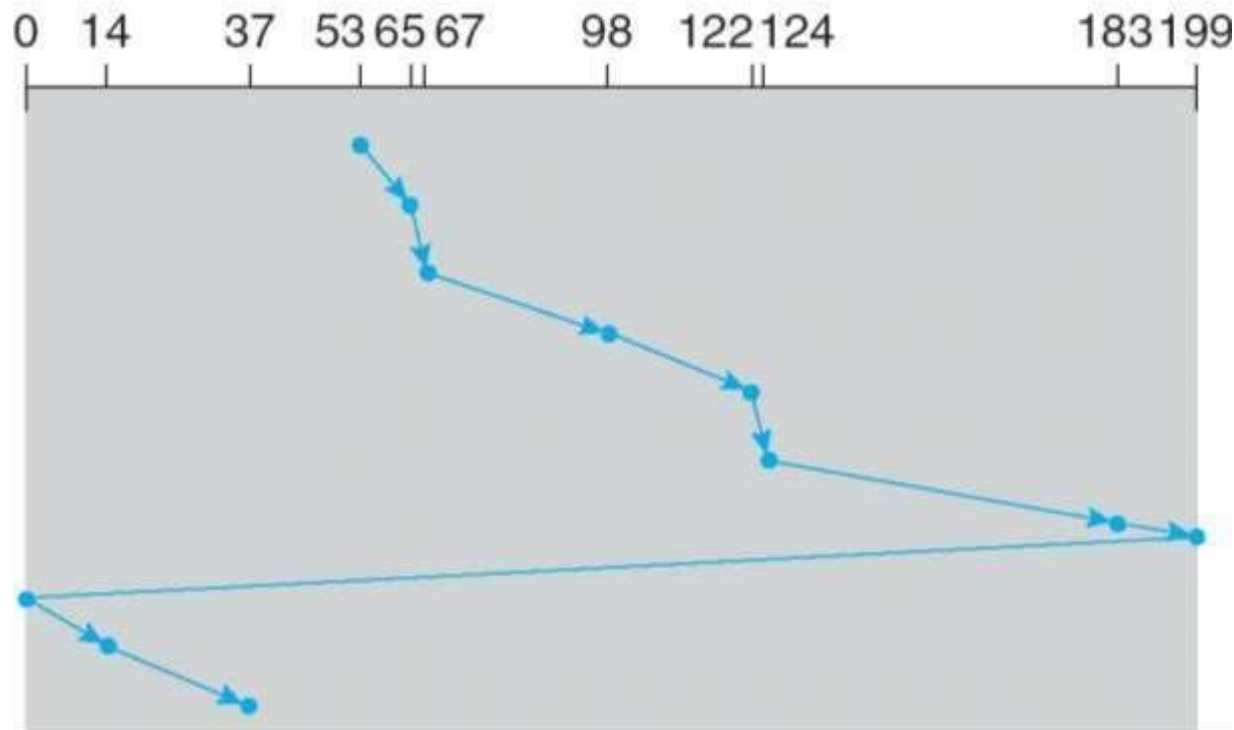


C-SCAN

- Provides a more uniform wait time than SCAN
- The head moves from one end of the disk to the other, servicing requests as it goes
 - When it reaches the other end, however, it immediately returns to the beginning of the disk, without servicing any requests on the return trip
- Treats the cylinders as a circular list that wraps around from the last cylinder to the first one
- Total number of cylinders?

C-SCAN (Cont.)

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53

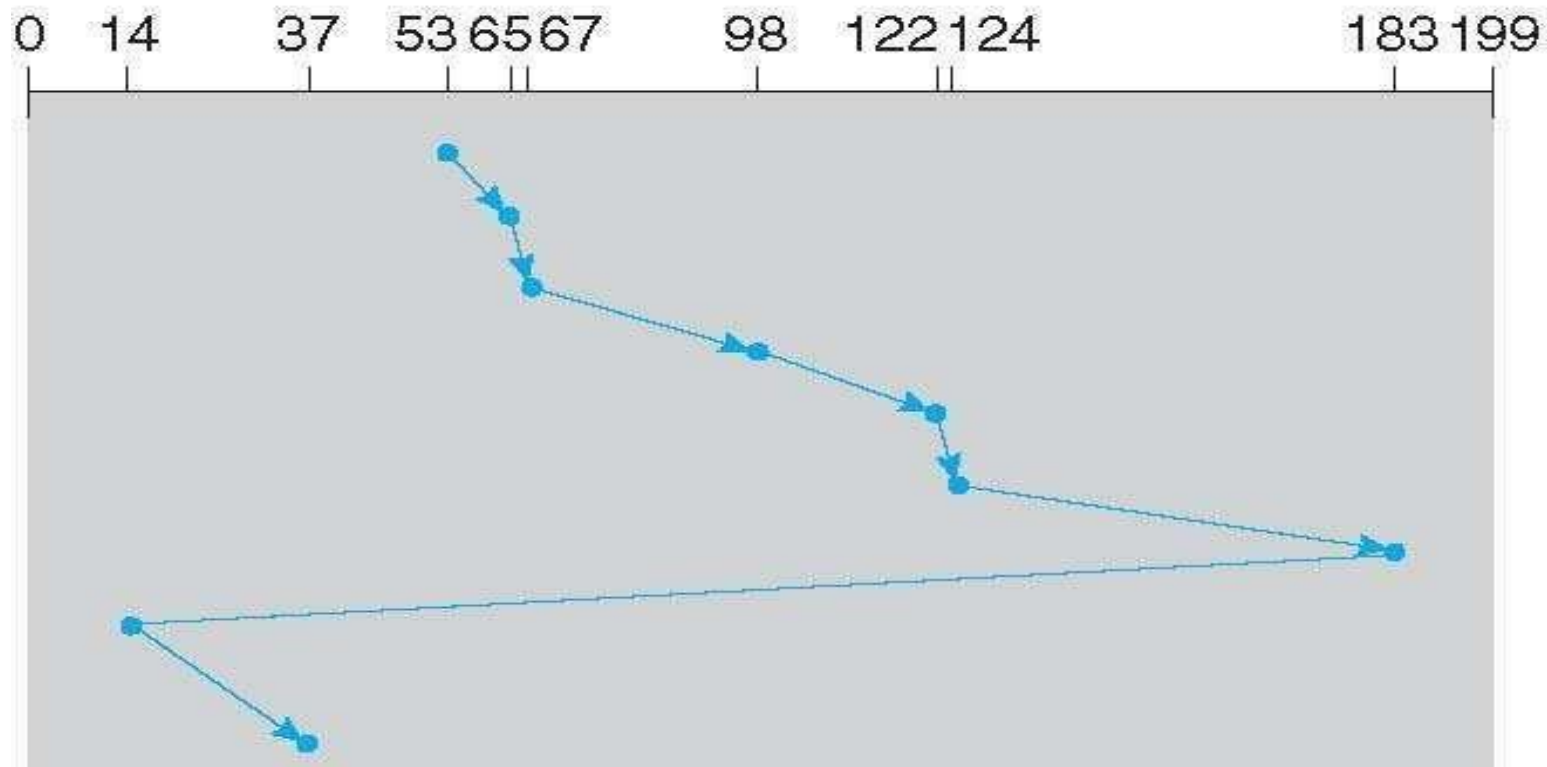


C-LOOK

- LOOK a version of SCAN, C-LOOK a version of C-SCAN
- Arm only goes as far as the last request in each direction, then reverses direction immediately, without first going all the way to the end of the disk
- Total number of cylinders?

C-LOOK (Cont.)

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53



Selecting a Disk-Scheduling Algorithm

- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk
 - Less starvation
- Performance depends on the number and types of requests
- Requests for disk service can be influenced by the file-allocation method
 - And metadata layout
- The disk-scheduling algorithm should be written as a separate module of the operating system, allowing it to be replaced with a different algorithm if necessary
- Either SSTF or LOOK is a reasonable choice for the default algorithm
- What about rotational latency?
 - Difficult for OS to calculate
- How does disk-based queueing effect OS queue ordering efforts?

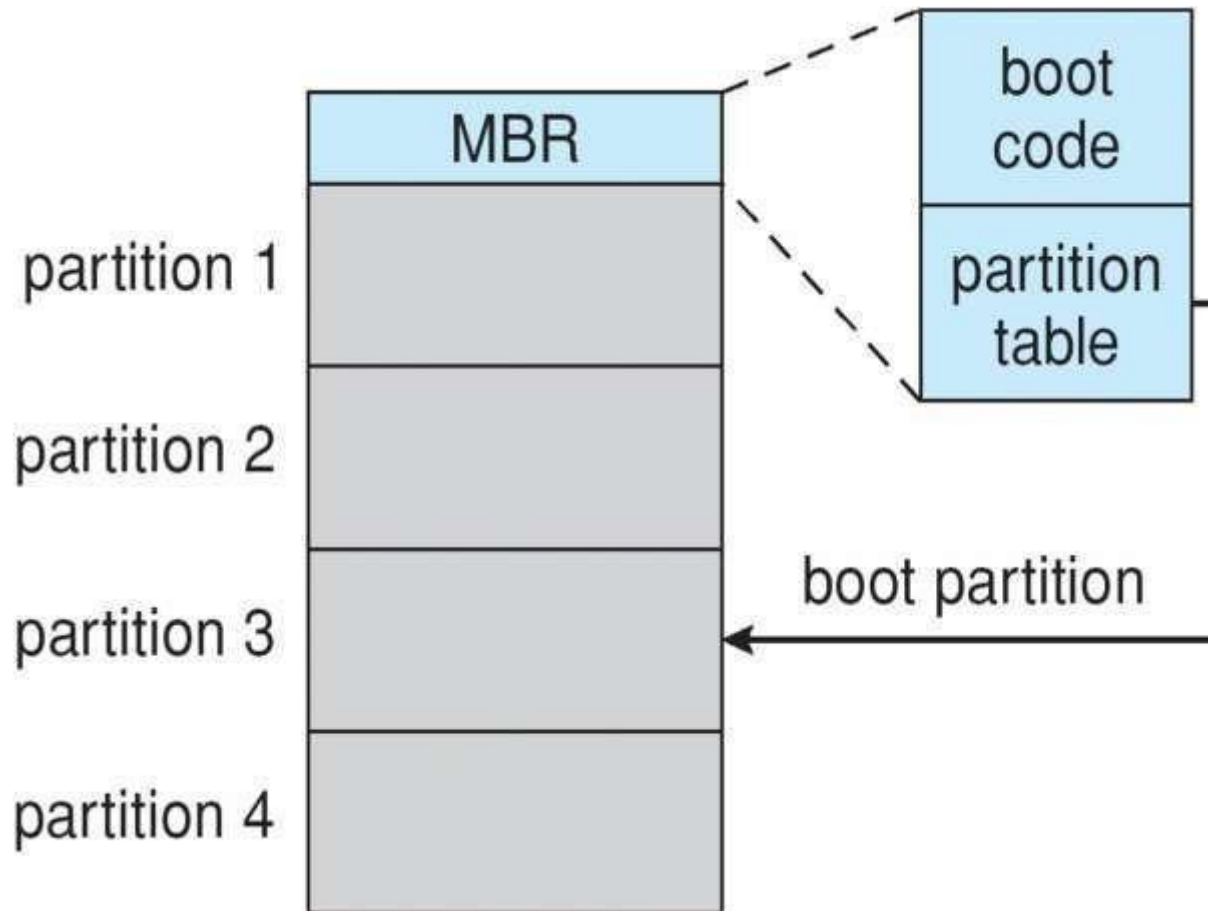
Disk Management

- **Low-level formatting**, or **physical formatting** — Dividing a disk into sectors that the disk controller can read and write
 - Each sector can hold header information, plus data, plus error correction code (**ECC**)
 - Usually 512 bytes of data but can be selectable
- To use a disk to hold files, the operating system still needs to record its own data structures on the disk
 - **Partition** the disk into one or more groups of cylinders, each treated as a logical disk
 - **Logical formatting** or “making a file system”
 - To increase efficiency most file systems group blocks into **clusters**
 - Disk I/O done in blocks
 - File I/O done in clusters

Disk Management (Cont.)

- Raw disk access for apps that want to do their own block management, keep OS out of the way (databases for example)
- Boot block initializes system
 - The bootstrap is stored in ROM
 - **Bootstrap loader** program stored in boot blocks of boot partition
- Methods such as **sector sparing** used to handle bad blocks

Booting from a Disk in Windows



Swap-Space Management

- Swap-space — Virtual memory uses disk space as an extension of main memory
 - Less common now due to memory capacity increases
- Swap-space can be carved out of the normal file system, or, more commonly, it can be in a separate disk partition (raw)
- Swap-space management
 - 4.3BSD allocates swap space when process starts; holds text segment (the program) and data segment
 - Kernel uses **swap maps** to track swap-space use
 - Solaris 2 allocates swap space only when a dirty page is forced out of physical memory, not when the virtual memory page is first created
 - File data written to swap space until write to file system requested
 - Other dirty pages go to swap space due to no other home
 - Text segment pages thrown out and reread from the file system as needed
- What if a system runs out of swap space?

Data Structures for Swapping on Linux Systems

