

CMSC 412

Spring 2007

Storage

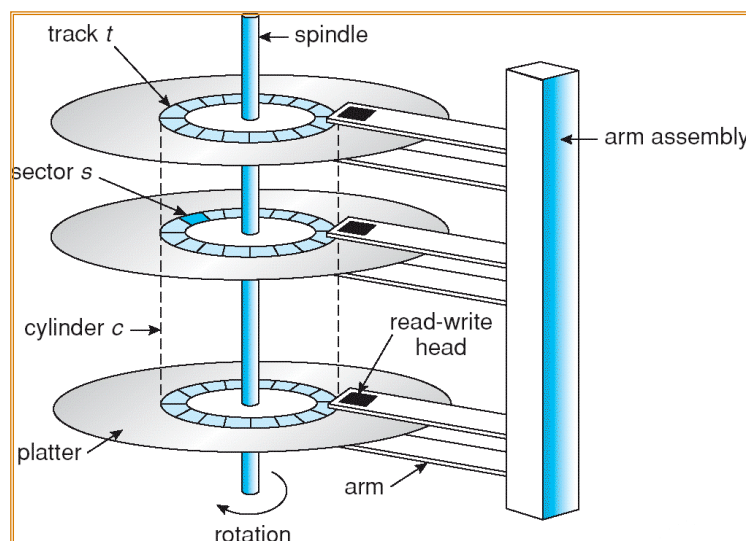
Announcements

- Reading
 - Chapter 12
- Office Hours
 - **No class Wednesday**, will have office hours in Linuxlab instead
- Project 4
 - Due Wednesday
- Project 5 (file system implementation)
 - Posted Wednesday

Magnetic Disks

- Provide bulk of secondary storage of modern computers
 - Drives rotate at 60 to 200 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**)

Moving-head Disk Mechanism



Magnetic Tapes

- Relatively permanent and holds large quantities of data
 - Mainly used for backup, storage of infrequently-used data, transfer medium between systems
 - We use 500 GB tapes in the Dept.
- Access time slow
 - Random access ~1000 times slower than disk
 - Once data under head, transfer rates like disk
- Price/Performance
 - 500 GB/\$100 vs. 250 GB disk/\$80
 - Cost of Dept. tape drive is prohibitive: \$10,000!

Disk Structure

- Disk drives are addressed as large one-dimensional arrays of *logical blocks*, 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

Performance: Disk Scheduling

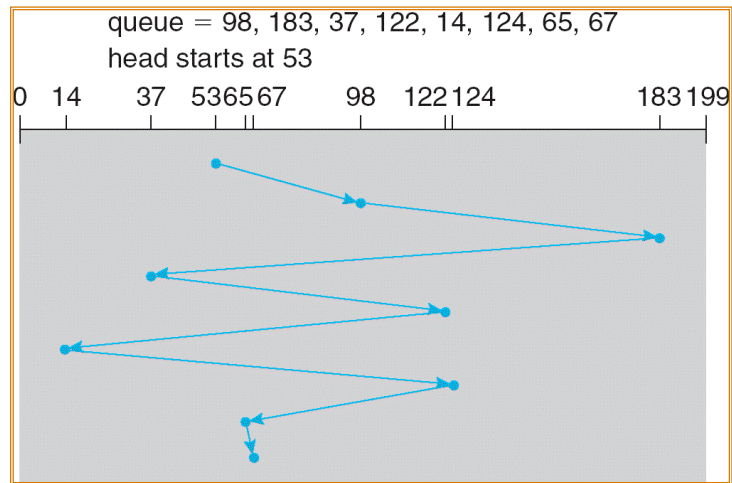
- **Access time** (*aka* Wait time)
 - Time to read the first block of a request. Units in ms.
- **Disk bandwidth**
 - Transfer rate averaged across all requests. Units in bytes/sec.
- **Goal: minimize access time and maximize bandwidth**
 - Core of approach: minimize seek time for each block of a request; seek time \approx seek distance

Disk Scheduling

- Several algorithms exist to schedule the servicing of disk I/O requests
 - Request consists of
 - Read/write
 - Disk address
 - Memory address
 - Number of sectors to transfer
- Illustrate scheduling using the following request queue of disk addresses (0-199):
98, 183, 37, 122, 14, 124, 65, 67
Head pointer 53

FCFS

(First-come First-served)

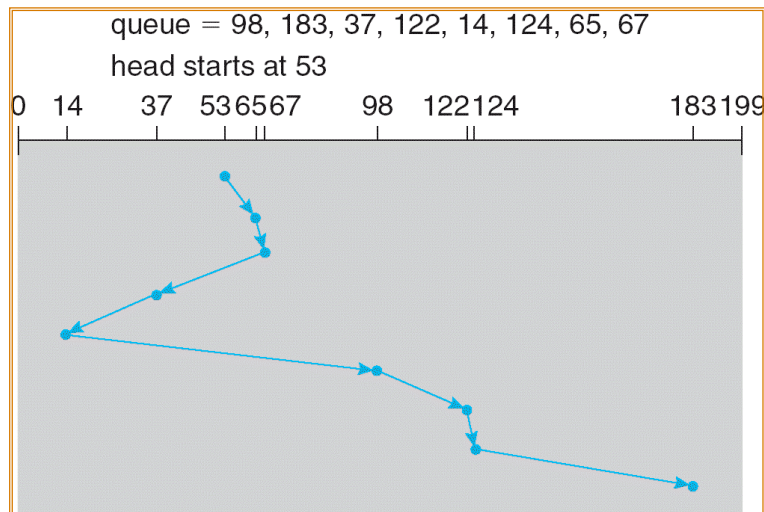


SSTF

(Shortest Seek Time First)

- Selects the request with the minimum seek time from the current head position
- SSTF scheduling is a form of shortest job first (SJF) scheduling
 - may cause starvation of some requests

SSTF example



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
- Sometimes called the *elevator algorithm*

SCAN example

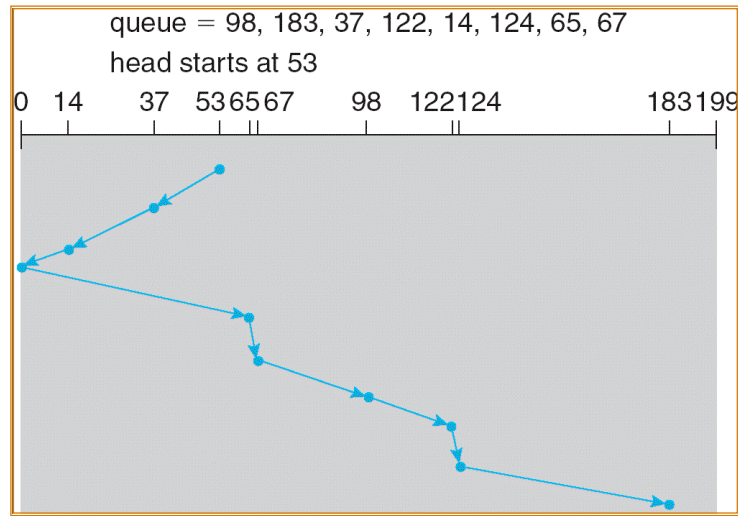
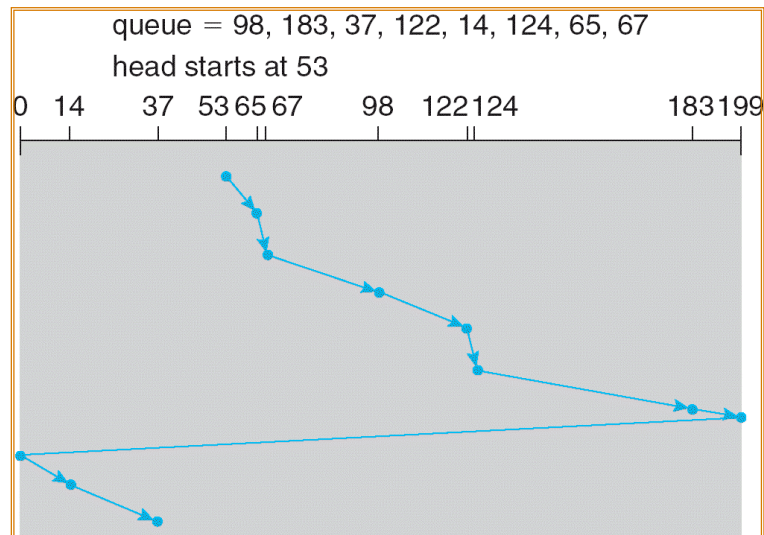


Illustration shows total head movement of 208 cylinders

C-SCAN

- The head moves from one end of the disk to the other servicing requests as it goes. When it reaches the other end 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
- Provides a more uniform wait time than SCAN

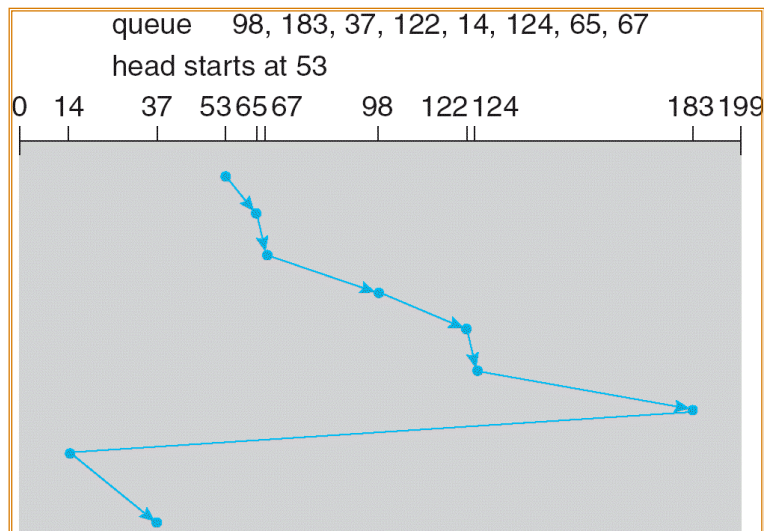
C-SCAN example



C-LOOK

- 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.

C-LOOK Example



Which should I use?

- SSTF is common
 - Maximizes bandwidth
- SCAN and C-SCAN perform better for systems that heavily load the disk
 - Less chance of starvation
- Performance depends on the number and types of requests
 - Which algorithm performs best when the request queue always has one element in it?

More tradeoffs

- Requests for disk service can be influenced by the file-allocation method
- The disk-scheduling algorithm should be written as a separate module of the operating system
 - Principle of abstraction
- Prioritization in the OS
 - Paging given higher priority?
 - Reads/writes for higher priority processes?

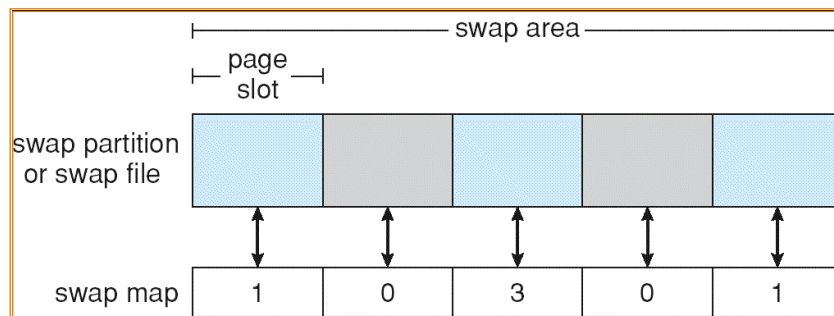
Bad Blocks

- Some sectors on the disk will go bad
 - Reported to the OS by the disk
- Recovery can be in the disk or the OS
 - MS-DOS FAT contains information about bad blocks
 - SCSI disks keep spare sectors and swap them in on reboot for bad addresses

Swap-Space Management

- Swap-space – Virtual memory uses disk space as an extension of main memory.
- Swap-space can be carved out of the normal file system, or, more commonly, it can be in a separate disk partition.
- 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 page is forced out of physical memory, not when the virtual memory page is first created
 - And only for non text-based (I.e., code) pages

Data Structures for Swapping on Linux Systems



Swap map contains the number of references to a given swapped page

RAID

- Redundant Arrays of Inexpensive Disks
- Compared to a single disk, can be used
 - To provide better performance
 - Via parallelism
 - Better reliability
 - Via redundancy
 - Or both
- Various standard RAID levels
 - Implement the above differently

Operating System Issues

- Major OS jobs are to manage physical devices and to present a virtual machine abstraction to applications
- For hard disks, the OS provides two abstractions:
 - Raw device - an array of data blocks.
 - File system - the OS queues and schedules the interleaved requests from several applications.

API for Tape Drives

- Tapes are presented as a raw device.
 - Usually the tape drive is reserved for the exclusive use of one application at a time.
- Since the OS does not provide file system services, the application must decide how to use the array of blocks.
 - Database management system
 - User-level file system

Tape Drive Operations

- **locate** positions the tape to a specific logical block (as opposed to seek).
- The **read position** op. reads the data.
- The **space** op. moves from current pos.
- Tape drives are “append-only” devices; updating a block in the middle of the tape also effectively erases everything beyond that block.
 - An EOT mark is placed after a block that is written.

File Naming

- The issue of naming files on removable media is especially difficult when we want to write data on a removable cartridge on one computer, and then use the cartridge in another computer.

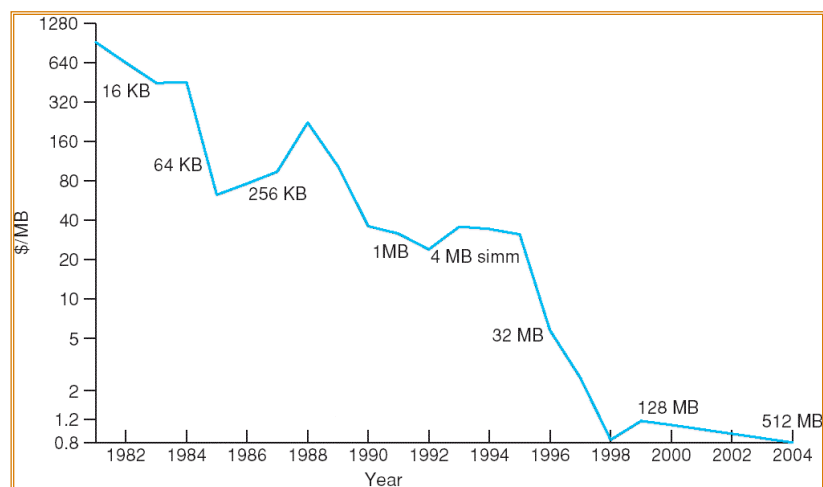
Reliability

- A fixed disk drive is likely to be more reliable than a removable disk or tape drive.
- An optical cartridge is likely to be more reliable than a magnetic disk or tape.
- A head crash in a fixed hard disk generally destroys the data, whereas the failure of a tape drive or optical disk drive often leaves the data cartridge unharmed.

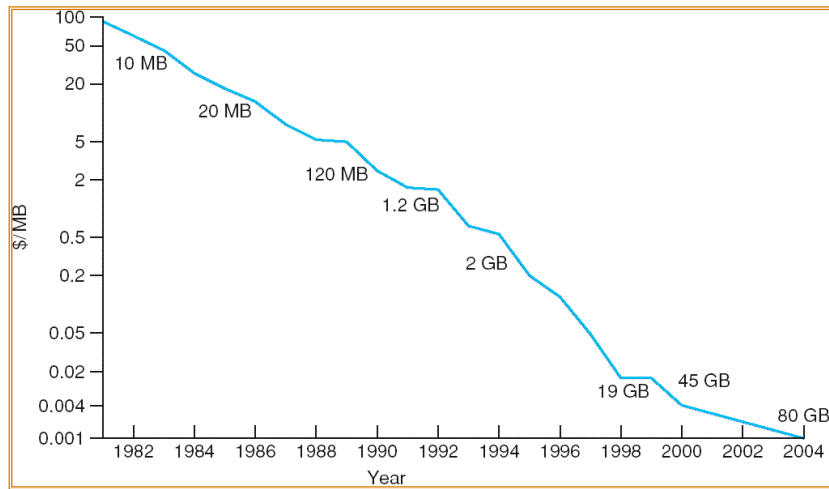
Cost

- Main memory is much more expensive than disk storage
- The cost per megabyte of hard disk storage is competitive with magnetic tape if only one tape is used per drive.
- The cheapest tape drives and the cheapest disk drives have had about the same storage capacity over the years.
- Tertiary storage gives a cost savings only when the number of cartridges is considerably larger than the number of drives.

Price of DRAM



Price of Magnetic Hard Disk



Price of Tape Drive

