Storage Systems

Disk Scheduling

Tertiary Storage Media

Emin Gun Sirer

Disk Scheduling

- Disk access time has two major components
  - *Seek time* is the time for the disk to move the heads to the cylinder containing the desired sector.
  - *Rotational delay* is the additional time waiting for the disk to rotate the desired sector to the disk head.
- Minimize seek time
- Seek time ? seek distance
- *Effective 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.

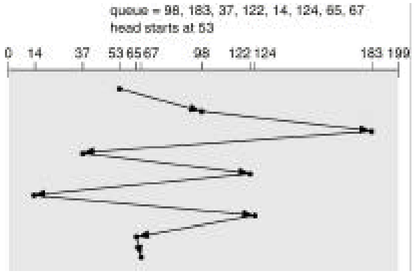
Disk Scheduling

- Disk controller is free to service outstanding requests in an order it chooses
  - Subject to constraints, e.g. inode writes after data writes
- Assume a disk with 200 (0-199) cylinders
  - Data read requests (no writes) on cylinders:

    98, 183, 37, 122, 14, 124, 65, 67
  - The disk head is at 53
- Simplest approach is to service the requests in the order of their arrival

FCFS

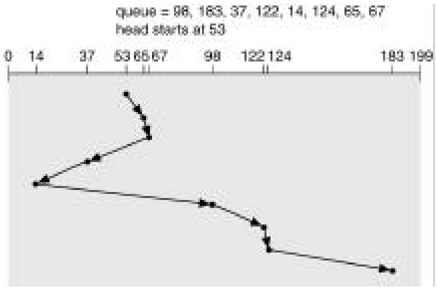
Total head movement of 640 cylinders.



SSTF

- 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.

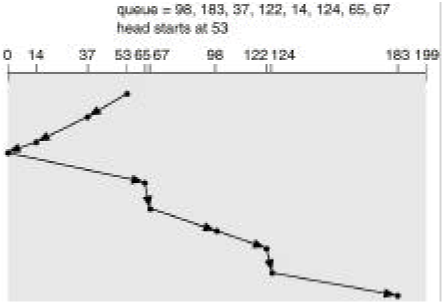
SSTF



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*.
- Illustration shows total head movement of 208 cylinders.

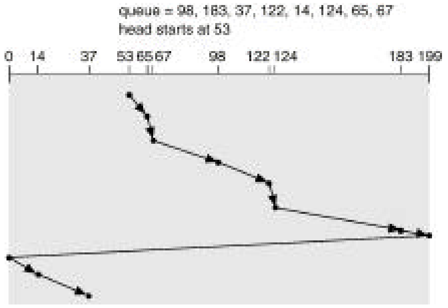
SCAN



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.

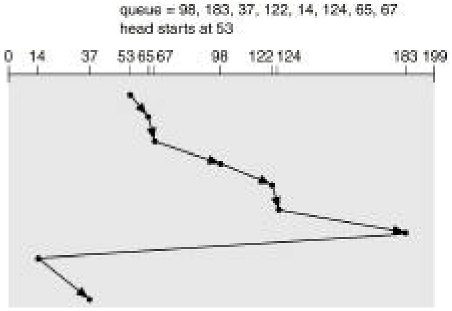
C-SCAN



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



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
- Performance depends on the number and types of requests
- Requests for disk service can be influenced by the file-allocation method
- Either SSTF or LOOK is a reasonable choice for the default algorithm

Tertiary Storage Devices

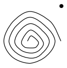
- Low cost is the defining characteristic of tertiary storage
  - Tradeoff between cost and access time
  - Tradeoff between data stability and access time
- Generally, tertiary storage is built using *removable media*
  - Floppy disks
  - ZIP drives
  - CD-ROMs
  - CD-RWs
  - DVDs
  - Magneto-optical storage
  - MEMS
  - Tapes

Removable Disks

- Floppy disk — thin flexible disk coated with magnetic material, enclosed in a protective plastic case
  - Most floppies hold about 1.4 MB, higher capacity (~GBs) is possible, standardization barriers
  - Removable magnetic disks can be nearly as fast as hard disks, but they are at a greater risk of damage from exposure
- Microdrive
  - Environmentally-sealed, miniature hard disk
  - ~1G capacity
  - Fast, reliable

CD-ROMs

- Aluminum encased in plastic
  - A laser light is bounced off the aluminum surface
  - **Pits** and **lands** on the aluminum bounce back light with different phase shifts
  - Pits stick out toward the laser ¼ wavelength
  - Pit/land and land/pit transitions encode 1s and 0s
- First mass-market digital media, launched in 1980
  - Designed to last 100 years



- Differs from magnetic disk drives in that the entire disk is one concentric spiral
  - Like a record
  - Possible to seek to locations on the spiral
- Low-level error correction
  - 7203 bytes allocated to carry a 2048 byte sector

CD-Rs and CD-RWs

- Like a CD-ROM, but recordable
  - An extra groove to guide the laser for writing
  - Markers on the groove enable precise rotational speed measurement
- CD-Recordable: Gold layer on top, with a dye layer below
  - Initially, the dye is transparent
  - Using the laser at high power alters the dye, reduces its reflectivity
  - Thus the dye simulates the pits and lands on CD-ROMs
- CD-RW: Silver, indium, antimony, tellurium alloy layer
  - Alloy has two stable states: crystalline and amorphous
  - High power laser turns it into amorphous, medium power turns it into crystalline, low power reads the reflectivity
- CD-Rs cannot be erased accidentally, CD-RW are more expensive

WORM Disks

- The data on read-write disks can be modified over and over.
- WORM (“Write Once, Read Many Times”) disks can be written only once
- Very durable and reliable
- Usually arranged in a jukebox
- Contrast with *read only* disks, such as CD-ROM and DVD, which come from the factory with the data pre-recorded

Magneto-Optical Disks

- A magneto-optic disk records data on a rigid platter coated with magnetic material
  - Laser heat is used to amplify a large, weak magnetic field to record a bit
  - Laser light is also used to read data (Kerr effect)
  - The magneto-optical head flies much farther from the disk surface than a magnetic disk head, and the magnetic material is covered with a protective layer of plastic or glass; resistant to head crashes
- Optical disks (CD-ROMs, CD-Rs, CD-RWs) do not use magnetism; they employ special materials that are altered by laser light

Tapes

- Compared to a disk, a tape is less expensive and holds more data, but random access is much slower.
- Tape is an economical medium for purposes that do not require fast random access, e.g., backup copies of disk data, holding huge volumes of data.
- Large tape installations typically use robotic tape changers that move tapes between tape drives and storage slots in a tape library.
  - stacker – library that holds a few tapes
  - silo – library that holds thousands of tapes
- A disk-resident file can be *archived* to tape for low cost storage; the computer can *stage* it back into disk storage for active use.

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.

Application Interface

- Most OSs handle removable disks almost exactly like fixed disks — a new cartridge is formatted and an empty file system is generated on the disk.
- Tapes are presented as a raw storage medium, i.e., an application does not open a file on the tape, it opens the whole tape drive as a raw device.
- Usually the tape drive is reserved for the exclusive use of that application.
- Since the OS does not provide file system services, the application must decide how to use the array of blocks.
- Since every application makes up its own rules for how to organize a tape, a tape full of data can generally only be used by the program that created it.

Tape Drives

- Tapes export different primitives than disks
  - Disks: seek, read, write
  - Tapes: locate, read, write, space
- **locate** positions the tape to a specific logical block, not an entire track
  - Corresponds to seek
- The **read position** operation returns the logical block number where the tape head is.
- The **space** operation enables relative motion
- 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.

Hierarchical Storage Management (HSM)

- A hierarchical storage system extends the storage hierarchy beyond primary memory and secondary storage to incorporate tertiary storage — usually implemented as a jukebox of tapes or removable disks.
- Usually incorporate tertiary storage by extending the file system.
  - Small and frequently used files remain on disk.
  - Large, old, inactive files are archived to the jukebox.
- HSM is usually found in supercomputing centers and other large installaitons that have enormous volumes of data.

Speed

- Two aspects of speed in tertiary storage are bandwidth and latency.
- Bandwidth is measured in bytes per second.
  - Sustained bandwidth – average data rate during a large transfer; # of bytes/transfer time.
    Data rate when the data stream is actually flowing.
  - Effective bandwidth – average over the entire I/O time, including **seek** or **locate**, and cartridge switching.
    Drive's overall data rate.

Speed

- Access latency – amount of time needed to locate data.
  - Access time for a disk: <20 milliseconds.
  - Access time for tape: ~1 minute
  - Generally say that random access within a tape cartridge is about a thousand times slower than random access on disk.
- The low cost of tertiary storage is a result of having many cheap cartridges share a few expensive drives.
- A removable library is best devoted to the storage of infrequently used data, because the library can only satisfy a relatively small number of I/O requests per hour.

Reliability

- Reliability is a function of different failure modes
  - Mechanical properties of the head-surface interface
  - Stability of the storage medium
  - Reusability of the storage medium
- A fixed disk drive is likely to be more reliable than a removable disk or tape drive
  - Environment is better controlled
- An optical cartridge is likely to be more reliable than a magnetic disk or tape
  - Head failure often leaves data intact
- Optical devices are likely to be more stable than magnetic devices
  - Tapes require constant rewinding and forwarding

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
- Tertiary storage gives a cost savings only when the number of cartridges is considerably larger than the number of drives
  - The OS may need schemes, like RAID or multilevel IO systems, in order to address failures, hide latencies, provide availability, and reduce storage costs