Chapter 14: Mass-Storage Systems

- Disk Structure
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Disk Attachment
- Tertiary Storage Devices

Disk Structure

- Disk drives is large array of logical blocks
- index is address
- logical block is the smallest unit of transfer
- 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

Disk Scheduling

- Goal: efficient use of hardware:
  - fast access time
  - high disk bandwidth
- Access time components
  - Seek time: move the heads to cylinder with desired sector
  - Rotational latency: time for the disk to rotate to desired sector
- Minimize seek time
  - Seek time = seek distance
- Disk bandwidth
  - bytes transferred divided by time from first request for service to completion of last transfer
Disk Scheduling (Cont.)

- Several algorithms exist to schedule the servicing of disk I/O requests
- We illustrate them with a request queue (0-199):
  
  98, 183, 37, 122, 14, 124, 65, 67

  Head pointer 53

---

FCFS

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

Total head movement: 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
### SSTF (Cont.)

- Head starts at 53
- Total head movement: 236 cylinders

---

### SCAN

- Disk arm starts at one end of the disk
- 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 (Cont.)

- Queue = 98, 183, 37, 122, 14, 124, 65, 67
- Head starts at 53
- Total head movement: 208 cylinders
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 cylinders as a circular list that wraps around from the last cylinder to the first one

C-SCAN (Cont.)

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
Selecting a Disk-Scheduling Algorithm

- SSTF is common and has natural appeal
- SCAN and C-SCAN perform better for heavy disk load
- Performance depends on number and type of requests
- Requests are influenced by file allocation method

- 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 reasonable choice as default algorithm

Disk Management

- **Low-level formatting, or physical formatting:**
  - Dividing a disk into sectors that the disk controller can read and write
  - OS needs to record its data structures on disk:
    - Partition the disk into one or more groups of cylinders
    - Logical formatting or “making a file system”
- Boot block initializes system:
  - The bootstrap is stored in ROM
  - Bootstrap loader program
- Bad blocks handled via sector sparing
**Swap-Space Management**

- **Swap-space**
  - Virtual memory uses disk space as extension of memory
- **Swap-space is either**
  - normal file system
  - 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.

**4.3 BSD Text-Segment Swap Map**

- Fixed size blocks
4.3 BSD Data-Segment Swap Map

variable size blocks

RAID Structure

- RAID
  - Redundant Array of Inexpensive Disks
  - multiple disk drives provide reliability via redundancy
- RAID is arranged into six different levels
  - Stripping uses a group of disks as one storage unit
  - Mirroring or shadowing keeps duplicate of each disk
  - Block interleaved parity uses much less redundancy

RAID Levels
Disk Attachment

- Disks may be attached one of two ways:
  1. Host attached via an I/O port
  2. Network attached via a network connection

Network-Attached Storage
Low cost is the defining characteristic of tertiary storage.

Generally, tertiary storage is built using removable media. Examples include:
- Floppy disks
- CD-ROM, CD-R, CD-RW
- DVD, DVD-R
- Tape drives
- USB drives
- ...
Removable Disks (Cont.)

- 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-optic 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 do not use magnetism; they employ special materials that are altered by laser light.

WORM Disks

- "Write Once, Read Many Times"

- Thin aluminum film sandwiched between two glass or plastic platters

- To write a bit, the drive uses a laser light to burn a small hole through the aluminum, information can be destroyed but not altered

- Very durable and reliable

- Read Only disks (CD-ROM and DVD) are factory pre-recorded.

Tapes

- Less expensive, holds more data, but slow random access

- Economical medium:
  - backup copies of disk data, holding huge volumes of data

- Available:
  - robotic tape changers that move tapes between tape drives and storage slots
    - stacker – library that holds a few tapes
    - silo – library that holds thousands of tapes
**Application Interface**

- Removable disks:
  - format a new cartridge is formatted
  - place empty file system on the disk
- Tapes
  - raw storage medium
  - application opens the whole tape drive as a raw device
  - tape drive is reserved for exclusive use of that application
- Application decides on how to use the array of blocks
- tape full of data can generally only be used by the program that created it

**Tape Drives**

- The basic operations for a tape drive differ from those of a disk drive
- **locate**
  - positions the tape to a specific logical block
  - not an entire track (corresponds to **seek**)
- **read position**
  - operation returns current logical block number
- **space** operation enables relative motion
- Tape drives are “append-only” devices, updating a block in the middle of the tape, erases everything beyond that block
- An EOT mark is placed after last block written

**File Naming**

- Few standards
  - CD
  - DVD
- others:
  - let application figure out how to access and interpret the data
Hierarchical Storage Management

- Storage hierarchy
  - primary, secondary, tertiary
- Tertiary storage via file system extensions
  - Small and frequently used files remain on disk
  - Large, old, inactive files are archived to tape jukebox
- Found in supercomputing centers and other large installations 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
    - includes seek or locate, and cartridge switching
    - Drive’s overall data rate

Speed (Cont.)

- Access latency – amount of time needed to locate data
  - Access time for a disk – move the arm to the selected cylinder and wait for the rotational latency;
    - < 35 milliseconds
  - Access on tape requires winding the tape reels until the selected block reaches the tape head;
    - tens or hundreds of seconds
- The low cost of tertiary storage is result of having many cheap cartridges share an expensive drive
Reliability

- Fixed disk drive is more reliable than 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 per Megabyte of Magnetic Hard Disk, From 1981 to 2000

Price per Megabyte of a Tape Drive, From 1984-2000