Chapter 12: Mass-Storage Systems

Objectives

- Describe the physical structure of secondary and tertiary storage devices and the resulting effects on the uses of the devices
- Explain the performance characteristics of mass-storage devices
- Discuss operating-system services provided for mass storage, including RAID and HSM

Mass Storage: Magnet disks

- **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**)
- **Head crash** results from disk head making contact with the disk surface
  - That’s bad

Moving-head Disk Mechanism

Estimate transferring rate

- Suppose that a disk drive spins at 7200 RPM, has a sector size of 512 bytes, and holds 160 sectors per track.
- What is sustained transfer rate of this drive in megabytes per second?
  - Disk spins 120 times per second
  - Each spin transfers a track of 80 KB (160x0.5K)
  - Sustained transfer rate is 120x80 = 9.6MB/s.
Consider seek time/rotation overhead

- 7200 RPM, sector size of 512 bytes, and 160 sectors per track.
- Average seek time for the drive is 8 milliseconds.
- Estimate # of random sector I/Os per second that can be done and the effective transfer rate for random-access of a sector.
  
  - Average rotational cost is time to travel half track: \( \frac{1}{120} \times 50\% = 4.167 \text{ms} \)
  
  - Transfer time is 8 ms to seek
    + 4.167 ms rotational latency
    + 0.052 ms (reading one sector takes 0.5MB/9.6MB).
    = 12.219 ms
  
  - # of random sector access/second = \( \frac{1}{0.012219} = 81.8 \)
  
  - Effective transferring rate: 0.5 KB/0.012219s = 0.0409 MB/s.

More fine-grain consideration

- 7200 RPM, sector size of 512 bytes, and 160 sectors per track.
- This drive has 7000 cylinders, 20 tracks per cylinder.
- Head switch time (from one platter to another) of 0.5 millisecond.
- An adjacent cylinder seek time of 2 milliseconds.
- What is the total time and sustained transfer rate for a huge transfer of 100 cylinders of data?
  
  - # bytes transferred: 100 cyl * 20 trk/cyl * 80KB/trk, i.e., 160,000 KB.
  
  - Transfer time: rotation time + track switch time + cylinder switch time.
    - Rotation time is 2000 trks/120 trks per sec = 16.667 s.
    - Track switch time is 19 switch per cyl * 100 cyl * 0.5 ms = 950 ms.
    - Cylinder switch time is 99 * 2 ms = 198 ms.
    - Total time is 16.667 + 0.950 + 0.198 = 17.815 s.
    - Thus the transfer rate is 160 MB/17.815 = 8.98 MB/s.

Solid State Drive (SSD)

- Relatively permanent and holds large quantities of data.
- Random access ~1000 times slower than disk.
- Mainly used for backup, storage of infrequently-used data, transfer medium between systems.
- Once data under head, transfer rates comparable to disk.
- 20-1.5TB typical storage.
- Common technologies are 4mm, 8mm, 19mm, LTO-2 and SDLT.

Disk Attachment

- Drive attached to computer via I/O bus.
- USB.
- SATA (replacing ATA, PATA, EIDE).
- SCSI.
  - Itself is a bus, up to 16 devices on one cable, SCSI initiator requests operation and SCSI targets perform tasks.
  
  - FC (Fiber Channel) 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.
  
  - Can be arbitrated loop (FC-AL) of 126 devices.

Magnetic Tape

- Relatively permanent and holds large quantities of data.
- Random access ~1000 times slower than disk.
- Mainly used for backup, storage of infrequently-used data, transfer medium between systems.
- Once data under head, transfer rates comparable to disk.
- 20-1.5TB typical storage.
- Common technologies are 4mm, 8mm, 19mm, LTO-2 and SDLT.
Network-Attached Storage

- Network-attached storage (NAS) is storage made available over a network rather than over a local connection (such as a bus)
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage
- New iSCSI protocol uses IP network to carry the SCSI protocol

Storage Area Network

- Common in large storage environments (and becoming more common)
- Multiple hosts attached to multiple storage arrays - flexible

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
- Access time has two major components
  - Seek time is the time for the disk are to move the heads to the cylinder containing the desired sector
  - Rotational latency is the additional time waiting for the disk to rotate the desired sector to the disk head
- 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

FCFS (First Come First Serve)

Illustration shows total head movement of 640 cylinders

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

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 208 cylinders

**C-SCAN (Circular-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-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 (Cont.)

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

0 14 37 53 65 67 98 122 124 183 199

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
- 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

Disk Management

- Low-level formatting, or physical formatting — Dividing a disk into sectors that the disk controller can read and write
- 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
  - 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
- Boot block initializes system
  - The bootstrap is stored in ROM
  - Bootstrap loader program
- Methods such as sector sparing used to handle bad blocks

Booting from a Disk in Windows 2000

- 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

Data Structures for Swapping on Linux Systems
RAID (Redundant Array of Inexpensive Disks)

- Multiple disk drives provide reliability via redundancy, increasing the mean time to failure.
- RAID is arranged into six different levels:
  - Hardware RAID with RAID controller vs software RAID.

RAID (Cont.)

- Several improvements in disk use techniques involve the use of multiple disks working cooperatively.
- Disk striping uses a group of disks as one storage unit.
- RAID schemes improve performance and improve the reliability of the storage system by storing redundant data:
  - Mirroring or shadowing (RAID 1) keeps a duplicate of each disk.
  - Striped mirrors (RAID 10) or mirrored stripes (RAID 0+1) provide high performance and high reliability.
  - Block interleaved parity (RAID 4, 5, 6) uses much less redundancy.
- RAID within a storage array can still fail if the array fails, so automatic replication of the data between arrays is common.
- Frequently, a small number of hot-spare disks are left unallocated, automatically replacing a failed disk and having data rebuilt onto them.

Raid Level 0

- Level 0 is nonredundant disk array.
- Files are striped across disks, no redundant info.
- High read throughput.
- Best write throughput (no redundant info to write).
- Any disk failure results in data loss.

Raid Level 1

- Mirrored Disks.
- Data is written to two places:
  - On failure, just use surviving disk.
  - On read, choose fastest to read.
  - Write performance is same as single drive, read performance is 2x better.
- Expensive.

RAID 5

- Block 1
- Block 2
- Block 3
- Block 4
- Block 5
- Block 6
- Block 7
- Block 8

- Error Check Blocks 1 and 2.
- Error Check Blocks 3 and 4.
- Error Check Blocks 5 and 6.
- Error Check Blocks 7 and 8.

- Logical Drive.
  - Physical Disk #1
  - Physical Disk #2
  - Physical Disk #3

6 RAID Levels

- RAID 0: non-redundant striping.
- RAID 1: mirrored data.
- RAID 2: memory-style error correcting codes.
- RAID 3: block-interleaved parity.
- RAID 4: clock-interleaved parity.
- RAID 5: block-interleaved distributed parity.
- RAID 6: P + Q redundancy.
**RAID (0 + 1) and (1 + 0)**

- Stripe on a set of disks
- Then mirror of data blocks is striped on the second set.

---

**Raid Level 0+1**

- When there are multiple files with different sizes.

---

**Raid Level 1+0**

- Pair mirrors first.
- Then stripe on a set of paired mirrors
- Better reliability than RAID 0+1

---

**Operating System Support**

- Major OS jobs are to manage physical devices and to present a virtual machine abstraction to applications

  - For hard disks, the OS provides two abstraction:
    - Raw device – an array of data blocks
    - File system – the OS queues and schedules the interleaved requests from several applications

---

**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
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.

Price per Megabyte of DRAM
From 1981 to 2004

Price per Megabyte of Magnetic Hard Disk
From 1981 to 2004

Price per Megabyte of a Tape Drive
From 1984-2000

End of Chapter 12