CS 333
Introduction to Operating Systems

Class 16 - Secondary Storage Management

Jonathan Walpole
Computer Science
Portland State University
Disk geometry

- Disk head, surfaces, tracks, sectors ...
# Comparison of disk technology

<table>
<thead>
<tr>
<th>Parameter</th>
<th>IBM 360-KB floppy disk</th>
<th>WD 18300 hard disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cylinders</td>
<td>40</td>
<td>10601</td>
</tr>
<tr>
<td>Tracks per cylinder</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Sectors per track</td>
<td>9</td>
<td>281 (avg)</td>
</tr>
<tr>
<td>Sectors per disk</td>
<td>720</td>
<td>35742000</td>
</tr>
<tr>
<td>Bytes per sector</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>Disk capacity</td>
<td>360 KB</td>
<td>18.3 GB</td>
</tr>
<tr>
<td>Seek time (adjacent cylinders)</td>
<td>6 msec</td>
<td>0.8 msec</td>
</tr>
<tr>
<td>Seek time (average case)</td>
<td>77 msec</td>
<td>6.9 msec</td>
</tr>
<tr>
<td>Rotation time</td>
<td>200 msec</td>
<td>8.33 msec</td>
</tr>
<tr>
<td>Motor stop/start time</td>
<td>250 msec</td>
<td>20 sec</td>
</tr>
<tr>
<td>Time to transfer 1 sector</td>
<td>22 msec</td>
<td>17 μsec</td>
</tr>
</tbody>
</table>
Disk zones

- **Constant rotation speed**
  - Want constant bit density

- **Inner tracks:**
  - Fewer sectors per track

- **Outer tracks:**
  - More sectors per track
Disk geometry

- **Physical Geometry**
  - The actual layout of sectors on the disk may be complicated
  - The controller does the translation
  - The CPU sees a “virtual geometry”.
Disk geometry

(192 sectors in each view)
Disk formatting

- A disk sector

<table>
<thead>
<tr>
<th>Preamble</th>
<th>Data</th>
<th>ECC</th>
</tr>
</thead>
</table>

- Typically
  - 512 bytes / sector
  - ECC = 16 bytes
Cylinder skew
Disk capacity

- **For communication...**
  - 1 Kbps = 1,000 bits per second \((10^3)\)
  - 1 Mbps = 1,000,000 bits per second \((10^6)\)
  - 1 Gbps = 1,000,000,000 bits per second \((10^9)\)

- **For disks and memories...**
  - \(K = \text{kilo} = 2^{10} = 1024\)
  - \(M = \text{mega} = 2^{20} = 1024 \times 1024 = 1,048,576\)
  - \(G = \text{giga} = 2^{30} = 1024^3 = 1,073,741,824\)
  - 1 GB = \(2^{30}\) bytes = 1,073,741,824 bytes
Sector interleaving

No Interleaving

Single Interleaving

Double Interleaving
Disk scheduling algorithms

- Time required to read or write a disk block determined by 3 factors
  - Seek time
  - Rotational delay
  - Actual transfer time

- Seek time dominates
  - Schedule disk heads to minimize it
Disk scheduling algorithms

- First-come first serve
- Shortest seek time first
- Scan → back and forth to ends of disk
- C-Scan → only one direction
- Look → back and forth to last request
- C-Look → only one direction
Shortest seek first (SSF)
Shortest seek first (SSF)

- Cuts arm motion in half
- Fatal problem:
  - Starvation is possible!
The elevator algorithm

- Use one bit to track which direction the arm is moving
  - Up
  - Down
- Keep moving in that direction
- Service the next pending request in that direction
- When there are no more requests in the current direction, reverse direction
The elevator algorithm

Diagram showing initial position and sequence of seeks over time.
Other disk scheduling algorithms

- First-come first serve
- Shortest seek time first
- Scan → back and forth to ends of disk
- C-Scan → only one direction
- Look → back and forth to last request
- C-Look → only one direction
Errors on disks

- Transient errors v. hard errors
- Manufacturing defects are unavoidable
  - Some will be masked with the ECC (error correcting code) in each sector
- Dealing with bad sectors
  - Allocate several spare sectors per track
- At the factory, some sectors are remapped to spares
  - Errors may also occur during the disk lifetime
- The sector must be remapped to a spare
  - By the OS
  - By the device controller
Using spare sectors

- **Spare sectors**
- **Bad sector**
- **Replacement sector**
- **Substituting a new sector**
- **Shifting sectors**
Handling bad sectors in the OS

- Add all bad sectors to a special file
  - The file is hidden; not in the file system
  - Users will never see the bad sectors
    - There is never an attempt to access the file

- Backups
  - Some backup programs copy entire tracks at a time
    - Efficient
  - Problem:
    - May try to copy every sector
    - Must be aware of bad sectors
Stable storage

- The model of possible errors:
  - Disk writes a block
  - If there is an error during a write...
    - It will probably be detected upon reading the block
  - Disk blocks can go bad spontaneously
    - But subsequent reads will detect the error
  - CPU can fail (just stops)
    - Disk writes in progress are detectable errors
  - Highly unlikely to lose the same block on two disks (on the same day)
Stable storage

- Use two disks for redundancy

- Each write is done twice
  - Each disk has N blocks.
  - Each disk contains exactly the same data.

- To read the data ...
  - you can read from either disk

- To perform a write ...
  - you must update the same block on both disks

- If one disk goes bad ...
  - You can recover from the other disk
Stable storage

- **Stable write**
  - Write block on disk #1
  - Read back to verify
  - If problems...
    - Try again several times to get the block written
    - Then declare the sector bad and remap the sector
    - Repeat until the write to disk #1 succeeds
  - Write same data to corresponding block on disk #2
    - Read back to verify
    - Retry until it also succeeds
Stable storage

- **Stable Read**
  - Read the block from disk #1
  - If problems...
    - Try again several times to get the block
  - If the block can not be read from disk #1...
    - Read the corresponding block from disk #2

- Our Assumption:
  - The same block will not simultaneously go bad on both disks
Stable storage

- **Crash Recovery**
- Scan both disks
- Compare corresponding blocks
- For each pair of blocks...
  - If both are good and have same data...
    - Do nothing; go on to next pair of blocks
  - If one is bad (failed ECC)...  
    - Copy the block from the good disk
  - If both are good, but contain different data...  
    - *(CPU must have crashed during a “Stable Write”)*
    - Copy the data from disk #1 to disk #2
Crashes during a stable write

(a)

Disk 1
Old

(b)

Disk 1
ECC error

(b)

Disk 1
Old

(c)

Disk 1
New

(d)

Disk 1
New

(e)

Disk 1
New
Stable storage

- Disk blocks can spontaneously decay
- **Given enough time...**
  - The same block on both disks may go bad
    - Data could be lost!
  - Must scan both disks to watch for bad blocks (e.g., every day)

- Many variants to improve performance
  - **Goal**: avoid scanning entire disk after a crash.
  - **Goal**: improve performance
    - Every stable write requires: 2 writes & 2 reads
    - Can do better...
RAID

- Redundant Array of Independent Disks
- Redundant Array of Inexpensive Disks

Goals:
- Increased reliability
- Increased performance
RAID

(d) Bit 1 Bit 2 Bit 3 Bit 4 Parity RAID level 3

(e) Strip 0 Strip 1 Strip 2 Strip 3 Strip 7 Strip 11 RAID level 4
  Strip 4 Strip 5 Strip 6 Strip 10
  Strip 9

(f) Strip 0 Strip 1 Strip 2 Strip 3 Strip 10 Strip 11 RAID level 5
  Strip 4 Strip 5 Strip 6
  Strip 7
  Strip 12
  P12-15 Strip 13
  P16-19 Strip 17
  Strip 16
  Strip 18
  Strip 19
CDs & CD-ROMs

- Spiral groove

2K block of user data

Pit
Land
CD-ROMs

- 32x CD-ROM = 5,000,000 Bytes/Sec
- SCSI-2 is twice as fast.
CD-R (CD-Recordable)

- Printed label
- Dark spot in the dye layer burned by laser when writing
- Protective lacquer
- Reflective gold layer
- Dye layer
- Polycarbonate
- Substrate
- Direction of motion
- Lens
- Prism
- Infrared laser diode
- Photodetector
Updating write-once media

- VTOC = Volume Table of Contents
- When writing, an entire track is written at once
- Each track has its own VTOC
Updating write-once media

- **VTOC** = Volume Table of Contents
- When writing, an entire track is written at once.
- Each track has its own VTOC.
- Upon inserting a CD-R,
  - Find the last track
  - Obtain the most recent VTOC
    - This can refer to data in earlier tracks
  - This tells which files are on the disk
  - Each VTOC supercedes the previous VTOC
Updating write-once media

- **VTOC = Volume Table of Contents**
- When writing, an entire track is written at once.
- Each track has its own VTOC.
- Upon inserting a CD-R,
  - Find the last track
  - Obtain the most recent VTOC
    - This can refer to data in earlier tracks
  - This tells which files are on the disk
  - Each VTOC supercedes the previous VTOC
- Deleting files?
CD-RW

- Uses a special alloy
- Alloy has two states, with different reflectivities
  - **Crystalline** (highly reflective) - Looks like “land”
  - **Amorphous** (low reflectivity) - Looks like a “pit”
- Laser has 3 powers
  - Low power: Sense the state without changing it
  - High power: Change to amorphous state
  - Medium power: Change to crystalline state
DVDs

- “Digital Versatile Disk”
  - Smaller Pits
  - Tighter Spiral
  - Laser with different frequency

- Transfer speed
  - 1X = 1.4MB/sec (about 10 times faster than CD)

- Capacity
  - 4.7 GB Single-sided, single-layer (7 times a CD-ROM)
  - 8.5 GB Single-sided, double-layer
  - 9.4 GB Double-sided, single-layer
  - 17 GB Double-sided, double-layer
DVDs

0.6mm Single-sided disk

0.6mm Single-sided disk

Polycarbonate substrate 1

Semireflective layer

Aluminum reflector

Adhesive layer

Aluminum reflector

Semireflective layer

Polycarbonate substrate 2