Magnetic Disk – common I/O device

Computer

<table>
<thead>
<tr>
<th>Processor (active)</th>
<th>Memory (passive) (where programs, data live when running)</th>
<th>Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control (“brain”)</td>
<td>Data Path (“brawn”)</td>
<td>Input</td>
</tr>
</tbody>
</table>

Devices

- Keyboard, Mouse
- Disk, Network
- Display, Printer
Magnetic Disk – common I/O device

- **A kind of Computer Memory:**
  - Information sorted by magnetizing ferrite material on surface of rotating disk (similar to tape recorder except binary signaling).

- **Nonvolatile storage** – retains its value without applying power to disk.

- **Types:**
  - *Floppy disks* – slower, less dense, removable.
  - *Hard Disk Drives* – faster, more dense, non-removable.

- **Purpose in computer systems (Hard Drive):**
  1. Long-term, inexpensive storage for files
  2. “Backup” for main-memory. Large, inexpensive, slow level in the memory hierarchy (discuss later)
The World's Smallest Hard Drive

**Hard disk**
The glass disk's metal coating—less than a thousandth of the thickness of a human hair—stores the same amount of data as a common DVD.

**Actuator**
Sweeps its microscopic read-and-write heads over both surfaces of the disk to position them for the transmission and retrieval of data.

**Locking latch**
The latch keeps the actuator from damaging the disk's surface if the unit is dropped.

**Spindle motor**
Powered by nine electromagnets, the motor spins the disk at 15 miles an hour.

**Rubber shock absorbers**
They help protect the unit from the frequent jar-ring and jostling suffered by portable devices.

**Circuit board**
The hard drive's brain, it directs all functions from disk speed to data flow.

It's bite-size, but it packs a huge byte.
A new inch-long hard disk drive made by Hitachi holds four gigabytes of data—about a thousand times the drive capacity of a desktop computer 20 years ago. It's the latest in a family of hard drives built to store data in handheld devices from PDAs to digital cameras. The hardest part of working small: Getting the actuator to move across the disk a mere 2,500,000th of an inch from its surface.

—Michael Kiesius
Hard Drives are Sealed. Why?

- The closer the head to the disk, the small the “spot size” and thus the denser the recording.
  - Measured in Gbit/in². ~60 is state of the art.

- Disks are sealed to keep the dust out.
  - Heads are designed to “fly” at around 5-20nm above the surface of the disk.
  - 99.999% of the head/arm weight is supported by the air bearing force (air cushion) developed between the disk and the head.
Photo of Disk Head, Arm, Actuator

- Spindle
- Arm
- Head
- Actuator
- Platters (12)
Several **platters**, with information recorded magnetically on both **surfaces** (usually)

Bits recorded in **tracks**, which in turn divided into **sectors** (e.g., 512 Bytes)

**Actuator** moves **head** (end of **arm**) over track (“**seek**”), wait for **sector** rotate under **head**, then read or write
Disk Device Performance

- **Disk Latency** = Seek Time + Rotation Time + Transfer Time + Controller Overhead

  - Seek Time? depends on no. tracks move arm, seek speed of disk
  - Rotation Time? depends on speed disk rotates, how far sector is from head
  - Transfer Time? depends on data rate (bandwidth) of disk \( f(\text{bit density}, \text{rpm}) \), size of request
Disk Device Performance

- Average distance of sector from head?
- 1/2 time of a rotation
  - 7200 Revolutions Per Minute ⇒ 120 Rev/sec
  - 1 revolution = 1/120 sec ⇒ 8.33 milliseconds
  - 1/2 rotation (revolution) ⇒ 4.17 ms

- Average no. tracks to move arm?
  - Disk industry standard benchmark:
    Sum all time for all possible seek distances from all possible tracks / # possible
    - Assumes average seek distance is random
Data Rate: Inner vs. Outer Tracks

• To keep things simple, originally same number of sectors per track
  • Since outer track longer, lower bits per inch

• Competition ⇒ decided to keep bits per inch (BPI) high for all tracks ("constant bit density")
  ⇒ More capacity per disk
  ⇒ More sectors per track towards edge
  ⇒ Since disk spins at constant speed, outer tracks have faster data rate

• Bandwidth outer track 1.7X inner track!
Disk Performance Model /Trends

• Capacity
  + 100%/year (2X / 1.0 yrs)

  Over time, this has grown so fast that # of platters has reduced (some even use only one now!)

• Transfer rate (BW)
  + 40%/year (2X / 2.0 yrs)

• Rotation + Seek time
  – 8%/year (1/2 in 10 yrs)

• MB/$
  > 100%/year (2X / 1.0 yrs)

  Fewer chips + bit density improvement

- 73.4 GB, 3.5 inch disk
- 10,000 RPM; 3 ms = 1/2 rotation
- 6 platters, 12 surfaces
- 13.2 Gbit/sq. in. areal den
- 10 watts (idle)
- 0.1 ms controller time
- 4.9 ms avg. seek
- 49 to 87 MB/s (internal)

Lower performance disks with similar capacities are available for $3-$9/GB

source: www.ibm.com
State of the Art: Barracuda 180 (~2001)

- 181.6 GB, 3.5-inch disk
- 7200 RPM; SCSI
  4.16 ms = 1/2 rotation
- 12 platters, 24 surfaces
- 31.2 Gbit/sq. in. areal den
- 10 watts (idle)
- 0.1 ms controller time
- 8.0 ms avg. seek
- 35 to 64 MB/s(internal)
- $7.50 / GB
  - (Lower capacity, ATA/IDE disks ~ $2 / GB)

source: www.seagate.com;

- 250 GB, 3.5 inch disk
- 7200 RPM;  
  4.17 ms = 1/2 rotation
- 3 platters, 6 surfaces
- 62 Gbit/sq. in. areal den
- 7.0 watts (idle)
- 0.1 ms controller time
- 8.5 ms avg. seek
- Transfer rate:  
  - 94.6 MB/s (internal)
  - 56.3 MB/s (external)
- $166, $1.50/GB

Lower capacities disks are available for < $1/GB

source: www.hitachi.com;
Disk Performance Example

• Calculate time to read 1 sector (512B) for Deskstar using advertised performance; sector is on outer track

Disk latency = average seek time + average rotational delay + transfer time + controller overhead

= 8.5 ms + 0.5 * 1/(7200 RPM) + 0.5 KB / (95 MB/s) + 0.1 ms

= 8.5 ms + 0.5 / (7200 RPM / (60000 ms/M)) + 0.5 KB / (95 KB/ms) + 0.1 ms

= 8.5 + 4.17 + 0.0048 + 0.1 ms = 12.77 ms

• How many CPU clock cycles is this?
Areal Density

- Bits recorded along a track
  - Metric is **Bits Per Inch** (BPI)
- Number of tracks per surface
  - Metric is **Tracks Per Inch** (TPI)
- We care about **bit density per unit area**
  - Metric is **Bits Per Square Inch**
  - Called **Areal Density**
- Areal Density = BPI x TPI
Early Disk History (IBM)

Data density
Mbit/sq. in.

Capacity of
Unit Shown
Megabytes

1973:
1.7 Mbit/sq. in
140 MBytes

1979:
7.7 Mbit/sq. in
2,300 MBytes

“Makers of disk drives crowd even more data into even smaller spaces”
Early Disk History

1989:
63 Mbit/sq. in
60,000 MBytes

1997:
1450 Mbit/sq. in
1600 MBytes

1997:
3090 Mbit/sq. in
8100 MBytes

“Makers of disk drives crowd even more data into even smaller spaces”
Areal Density

• Areal Density = BPI x TPI

• Change slope 30%/yr to 60%/yr about 1990

“Giant Magnetoresistive” effect head (highly sensitive)
Historical Perspective

• *Form factor* and *capacity* are more important in the marketplace than is performance

• Form factor evolution:

  1970s: Mainframes ⇒ 14 inch diameter disks

  1980s: Minicomputers, Servers
          ⇒ 8”, 5.25” diameter disks

  Late 1980s/Early 1990s:
  • PCs ⇒ 3.5 inch diameter disks
  • Laptops, notebooks ⇒ 2.5 inch disks
  • Palmtops didn’t use disks, so 1.8 inch diameter disks didn’t make it
1 inch disk drive!

- **2000 IBM MicroDrive:**
  - 1.7” x 1.4” x 0.2”
  - 1 GB, 3600 RPM, 5 MB/s, 15 ms seek
  - Digital camera, PalmPC?

- **2006 MicroDrive?**
  - 9 GB, 50 MB/s!
    Assuming it finds a niche in a successful product
    Assuming past trends continue
Suppose a typical book is 400 pages long, each page on average it has 40 lines of text, and there is an average of line of text has 80 characters. The PDF version of a chapter of our textbook takes 632 KB for 80 pages. The highest capacity 3.5-inch drive today is 250 GB.

1. How many books fit in such a disk today in text? in PDF? How long would it take to access a book from disk?

2. The Kreege Engineering library has 160,000 books and the Bancroft library has 400,000. If disk capacity doubles every year, when would a single disk contain all the books in these two libraries using text? PDF? Might this have impact on the campus? How?
Bonus slides
Fallacy: Use Data Sheet “Average Seek” Time

- Manufacturers needed standard for fair comparison (“benchmark”)
  - Calculate all seeks from all tracks, divide by number of seeks => “average”
- Real average would be based on how data is laid out on disk, and in what order it is accessed by application
  - Usually, tend to seek to tracks nearby, not to random track
- Rule of Thumb: observed average seek time is typically about 1/4 to 1/3 of quoted seek time (i.e., 3X-4X faster)
  - Deskstar 250 avg. seek: 8.5 ms ⇒ 2.83 ms
Fallacy: Use Data Sheet Transfer Rate

- Manufacturers quote the speed of the data rate off the surface of the disk.

- Sectors contain an error detection and correction field (can be 20% of sector size) plus sector number as well as data.

- There are gaps between sectors on track.

- Rule of Thumb: disks deliver about ~ 1/2-3/4 of internal media rate for data.

- For example, Deskstar 250 quotes 95 MB/s (max) internal media rate v. 56 MB/s (max) external data rate (60%)
Bonus: Future Disk Size and Performance

- Continued advance in capacity (60%/yr) and bandwidth (40%/yr)
- Slow improvement in seek, rotation (8%/yr)
- Time to read whole disk

<table>
<thead>
<tr>
<th>Year</th>
<th>Sequentially</th>
<th>Randomly (1 sector/seek)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>4 minutes</td>
<td>6 hours</td>
</tr>
<tr>
<td>2000</td>
<td>10 minutes</td>
<td>4 days</td>
</tr>
</tbody>
</table>

- 3.5” form factor make sense in 5-7 yrs?
Use Arrays of Small Disks?

- Katz and Patterson asked in 1987: Can smaller disks be used to close gap in performance between disks and CPUs?

**Conventional:**
- 4 disk designs
  - 3.5”
  - 5.25”
  - 10”
  - 14”

**Disk Array:**
- 1 disk design
  - 3.5”

**Low End** → **High End**
Replace Small Number of Large Disks with Large Number of Small Disks! (1988 Disks)

<table>
<thead>
<tr>
<th></th>
<th>IBM 3390K</th>
<th>IBM 3.5&quot; 0061</th>
<th>x70</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Capacity</strong></td>
<td>20 GBytes</td>
<td>320 MBytes</td>
<td>23 GBytes</td>
</tr>
<tr>
<td><strong>Volume</strong></td>
<td>97 cu. ft.</td>
<td>0.1 cu. ft.</td>
<td>11 cu. ft.</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>3 KW</td>
<td>11 W</td>
<td>1 KW</td>
</tr>
<tr>
<td><strong>Data Rate</strong></td>
<td>15 MB/s</td>
<td>1.5 MB/s</td>
<td>120 MB/s</td>
</tr>
<tr>
<td><strong>I/O Rate</strong></td>
<td>600 I/Os/s</td>
<td>55 I/Os/s</td>
<td>3900 I/Os/s</td>
</tr>
<tr>
<td><strong>MTTF</strong></td>
<td>250 KHrs</td>
<td>50 KHrs</td>
<td>??? Hrs</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>$250K</td>
<td>$2K</td>
<td>$150K</td>
</tr>
</tbody>
</table>

Disk Arrays potentially high performance, high MB per cu. ft., high MB per KW, but what about reliability?
Array Reliability

- **Reliability** - whether or not a component has failed
  - measured as Mean Time To Failure (MTTF)

- Reliability of N disks
  = Reliability of 1 Disk ÷ N
  (assuming failures independent)
  - 50,000 Hours ÷ 70 disks = 700 hour

- Disk system MTTF:
  Drops from 6 years to 1 month!

- Disk arrays too unreliable to be useful!
Redundant Arrays of (Inexpensive) Disks

- Files are "striped" across multiple disks

- Redundancy yields high data availability
  - **Availability**: service still provided to user, even if some components failed
    When disk fails, contents reconstructed from data redundantly stored in the array
  - ⇒ Capacity penalty to store redundant info
  - ⇒ Bandwidth penalty to update redundant info
Bonus slide: Berkeley History, RAID-I

- **RAID-I (1989)**
  - Consisted of a Sun 4/280 workstation with 128 MB of DRAM, four dual-string SCSI controllers, 28 5.25-inch SCSI disks and specialized disk striping software

- Today RAID is $27 billion dollar industry, 80% nonPC disks sold in RAIDs
Things to Remember

• Magnetic Disks continue rapid advance: 60%/yr capacity, 40%/yr bandwidth, slow on seek, rotation improvements, MB/$ improving 100%/yr.
  • Designs to fit high volume form factor
  • Quoted seek times too conservative, data rates too optimistic for use in system

• RAID
  • Higher performance with more disk arms per $
  • Adds availability option for small number of extra disks
  • Today RAID is $27 billion dollar industry, 80% nonPC disks sold in RAID; started at Cal