Magnetic Disk – common I/O device

- A kind of computer memory
  - Information stored by magnetizing ferrous material on surface of rotating disk
  - Similar to tape recorder except digital rather than analog data
- Nonvolatile storage
  - Retains its value without applying power to disk.
- Two Types
  - Floppy disks – slower, less dense, removable.
  - Hard Disk Drives (HDD) – faster, more dense, non-removable.
- Purpose in computer systems (Hard Drive):
  - Long-term, inexpensive storage for files
  - "Backup" for main-memory. Large, inexpensive, slow level in the memory hierarchy (virtual memory)

Disk Device Terminology

- 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 (1/2)

- Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead
  - Seek Time? depends on no. tracks to move arm, speed of actuator
  - Rotation Time? depends on speed disk rotates, how far sector is from head
  - Transfer Time? depends on data rate (bandwidth) of disk (bit density/rpm), size of request
Disk Device Performance (2/2)

- 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
- Size of Disk cache can strongly affect perf!
  - Cache built into disk system, OS knows nothing

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, given as fact that if it’s platters has reduced home even use only 1 hour!
- Transfer rate (BW): + 40%/yr (2X / 2 yrs)
- Rotation+Seek time: - 8%/yr (1/2 in 10 yrs)
- Areal Density
  - Bits recorded along a track: Bits/inch (BPI)
  - # of tracks per surface: Tracks/Inch (TPI)
  - We care about bit density per unit area: bits/inch²
  - Called Areal Density = BPI x TPI
  - ~120 Gb/In² is longitudinal limit
  - ~230 Gb/In² now with perpendicular
- GB/y: + 100%/yr (2X / 1.0 yrs)
  - Fewer chips = areal density

State of the Art: Two camps (2008)

- Performance
  - Enterprise apps, servers
  - E.g., Seagate Cheetah 15K.6
    - 3 Gb/s, Serial Attached SCSI, Fibre Channel
    - 450 GB, 3.5-inch disk
    - 4 disks, 8 heads
    - 15,000 RPM
    - 12-17 watts (idle-normal)
    - 3.4 ms avg. seek
    - 164 MB/s transfer rate
    - 1.6 Million Hrs MTBF
    - 5 year warranty
    - $1000 = $330 / GB
  - These use Perpendicular Magnetic Recording (PMR)!!!
  - Source: www.seagate.com

- Capacity
  - Mainstream, home uses
  - E.g., Seagate Barracuda 7200.11
    - SATA 3Gb/s NCO
    - SATA 1.5Gb/s NCO
    - 1 TB, 3.5-inch disk
    - 4 disks, 8 heads
    - 7,200 RPM
    - 8-12 watts (idle-normal)
    - ?? ms avg. seek (last one 8.5)
    - 115 MB/s transfer rate
    - 0.75 Million Hrs MTBF
    - 5 year warranty
    - $200 = 50 20 / GB

1 inch disk drive!

- Hitachi 2007 release
  - Development driven by
    - iPods & digital cameras
    - 8GB, 5-10MB/s (higher?)
    - 42.8 x 36.4 x 5 mm
- Perpendicular Magnetic Recording (PMR)
  - FUNDAMENTAL new technique
  - Evolution from Longitudinal
    - Starting to hit physical limit due to superparamagnetism
  - They say 10x improvement

Where does Flash memory come in?

- Microdrives and Flash memory (e.g., CompactFlash) are going head-to-head
  - Both non-volatile (no power, data ok)
  - Flash benefits: durable & lower power
    - (no moving parts, need to spin drives up/down)
  - Flash limitations: finite number of write cycles (wear on the insulating oxide layer around the charge storage mechanism)
- How does Flash memory work?
  - NMOS transistor with an additional conductor between gate and source/drain which “traps” electrons. The presence/absence is a 1 or 0.

Source: en.wikipedia.org/wiki/Flash_memory
What does Apple put in its iPods?

- Toshiba flash 1, 2GB
- Samsung flash 4, 8GB
- Toshiba flash 8, 16, 32GB
- Toshiba 1.8-inch HDD 80, 160GB

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”

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!

Replace Small # of Large Disks with Large # of Small!

<table>
<thead>
<tr>
<th>(1988 Disks)</th>
<th>IBM 3390K</th>
<th>IBM 3.5” 0061</th>
<th>x70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>20 GBytes</td>
<td>320 MBytes</td>
<td>23 GBytes</td>
</tr>
<tr>
<td>Volume</td>
<td>97 cu. ft.</td>
<td>0.1 cu. ft.</td>
<td>11 cu. ft.</td>
</tr>
<tr>
<td>Power</td>
<td>3 KW</td>
<td>11 W</td>
<td>1 KW</td>
</tr>
<tr>
<td>Data Rate</td>
<td>15 MB/s</td>
<td>1.5 MB/s</td>
<td>120 MB/s</td>
</tr>
<tr>
<td>I/O Rate</td>
<td>600 I/O/s</td>
<td>55 I/O/s</td>
<td>3900 I/O/s</td>
</tr>
<tr>
<td>MTTF</td>
<td>250 Khrs</td>
<td>50 Khrs</td>
<td>??? Khrs</td>
</tr>
<tr>
<td>Cost</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?

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
- Disks will still fail
- Contents reconstructed from data redundantly stored in the array
  - Capacity penalty to store redundant info
  - Bandwidth penalty to update redundant info
Berkeley History, RAID-I

- **RAID-1 (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 > tens billion dollar industry, 80% nonPC disks sold in RAIDs

**“RAID 0”: No redundancy = “AID”**

- Assume have 4 disks of data for this example, organized in blocks
- Large accesses faster since transfer from several disks at once

This and next 5 slides from RAID.edu http://www.acnc.com/04_01_00.html http://www.raid.com/04_00.html also has a great tutorial

**RAID 1: Mirror data**

- Each disk is fully duplicated onto its “mirror”
  - Very high availability can be achieved
- Bandwidth reduced on write:
  - 1 logical write = 2 physical writes
- Most expensive solution: 100% capacity overhead

**RAID 3: Parity**

- Parity computed across group to protect against hard disk failures, stored in P disk
- Logically, a single high capacity, high transfer rate disk
- 25% capacity cost for parity in this example vs. 100% for RAID 1 (5 disks vs. 8 disks)

Inspiration for RAID 5 (RAID 4 block-stripping)

- Small writes (write to one disk):
  - Option 1: read other data disks, create new sum and write to Parity Disk (access all disks)
  - Option 2: since P has old sum, compare old data to new data, add the difference to P
  - 1 logical write = 2 physical reads + 2 physical writes to 2 disks
- Parity Disk is bottleneck for Small writes: Write to A0, B1  both write to P disk

**RAID 5: Rotated Parity, faster small writes**

- Independent writes possible because of interleaved parity
  - Example: write to A0, B1 uses disks 0, 1, 4, 5, so can proceed in parallel
  - Still 1 small write = 4 physical disk accesses

en.wikipedia.org/wiki/Redundant_array_of_independent_disks
Digitally recorded information is stored on magnetic disks.

**Peer Instruction**

1. RAID 1 (mirror) and 5 (rotated parity) help with performance and availability.
2. RAID 1 has higher cost than RAID 5.
3. Small writes on RAID 5 are slower than on RAID 1.

---

**“And in conclusion…”**

- Magnetic Disks continue rapid advance: 60%/yr capacity, 40%/yr bandwidth, slow seek, rotation improvements, MB/s improving 100%/yr.
  - Designs to fit high volume form factor
  - PMR a fundamental new technology
  - Breaks through barrier
- RAID
  - Higher performance with more disk arms per $T$
  - Adds option for small # of extra disks
  - Can nest RAID levels
  - Today RAID is > tens-billion dollar industry.
  - 80% nonPC disks sold in RAID6s, started at Cal.

---

**Bonus slides**

- These are extra slides that used to be included in lecture notes, but have been moved to this, the “bonus” area to serve as a supplement.
- The slides will appear in the order they would have in the normal presentation.

---

**BONUS : Hard Drives are Sealed. Why?**

- The closer the head to the disk, the smaller the “spot size” and thus the denser the recording.
  - Measured in Gbit/in$^2$. >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.
  - >99.99% of the head/arm weight is supported by the air bearing force (or cushion) developed between the disk and the head.

---

**Historical Perspective**

- Form factor and capacity are more important in the marketplace than 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
  - Early 2000s:
    - MP3 players ⇒ 1 inch disks

---

The closer the head to the disk, the smaller the “spot size” and thus the denser the recording.
### Early Disk History (IBM)

<table>
<thead>
<tr>
<th>Year</th>
<th>Model 3340</th>
<th>Model 3370</th>
</tr>
</thead>
<tbody>
<tr>
<td>1973</td>
<td>1.7 Mbit/sq. in</td>
<td>7.7 Mbit/sq. in</td>
</tr>
<tr>
<td></td>
<td>140 MBytes</td>
<td>2,300 MBytes</td>
</tr>
</tbody>
</table>


### Early Disk History

<table>
<thead>
<tr>
<th>Year</th>
<th>1989</th>
<th>1997</th>
<th>1997:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>63 Mbit/sq. in</td>
<td>1450 Mbit/sq. in</td>
<td>3090 Mbit/sq. in</td>
</tr>
<tr>
<td></td>
<td>60,000 MBytes</td>
<td>1600 MBytes</td>
<td>8100 MBytes</td>
</tr>
</tbody>
</table>


### 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

  \[
  = 8.5 \text{ ms} + 0.5 \times \frac{1}{7200 \text{ RPM}} + 0.5 \text{ KB/100 KB/s} + 0.1 \text{ ms} \\
  = 8.5 \text{ ms} + 0.5 \times \frac{1}{7200 \text{ RPM}/(60000 \text{ ms/MB})} + 0.5 \text{ KB/100 KB/s} + 0.1 \text{ ms} \\
  = 8.5 + 4.17 + 0.005 + 0.1 \text{ ms} = 12.77 \text{ ms}
  \]

- How many CPU clock cycles is this?