iPod future: All the TV ⇒
Google’s VP believes that the “plummetsing price of storage” will allow iPods of the future (12 years) to hold “any video ever produced”. What he fails to note is how fast videos are being generated! -Dan
Magnetic Disk – common I/O device

Computer

Processor (active)

Control (“brain”)

Datapath (“brawn”)

Memory (passive) (where programs, data live when running)

Devices

Input

Output

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 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)
Photo of Disk Head, Arm, Actuator

- Spindle
- Arm
- Head
- Actuator
- Platters (1-12)
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 ($f(\text{bit density, rpm})$), 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
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 \((2X / 1.0 \text{ yrs})\)
  Over time, grown so fast that # of platters has reduced (some even use only 1 now!)

- Transfer rate (BW): + 40%/yr \((2X / 2 \text{ yrs})\)

- Rotation+Seek time: – 8%/yr \((1/2 \text{ in } 10 \text{ yrs})\)

- Areal Density
  - Bits recorded along a track: Bits/Inch \((\text{BPI})\)
  - # of tracks per surface: Tracks/Inch \((\text{TPI})\)
  - We care about bit density per unit area Bits/Inch\(^2\)
  - Called Areal Density = BPI x TPI
  - “~120 Gb/In\(^2\) is longitudinal limit”
  - “230 Gb/In\(^2\) now with perpendicular”

- GB/$: > 100%/year \((2X / 1.0 \text{ yrs})\)
  - Fewer chips + areal density

---

CS61C L37 Disks (9)
State of the Art: Two camps (2006)

**Performance**
- Enterprise apps, servers
  - E.g., Seagate Cheetah 15K.5
    - Ultra320 SCSI, 3 Gbit/sec, Serial Attached SCSI (SAS), 4Gbit/sec Fibre Channel (FC)
    - 300 GB, 3.5-inch disk
    - 15,000 RPM
    - 13 watts (idle)
    - 3.5 ms avg. seek
    - 125 MB/s transfer rate
    - 5 year warrantee
    - $1000 = $3.30 / GB

**Capacity**
- Mainstream, home uses
  - E.g., Seagate Barracuda 7200.10
    - Serial ATA 3Gb/s (SATA/300), Serial ATA 1.5Gb/s (SATA/150), Ultra ATA/100
    - 750 GB, 3.5-inch disk
    - 7,200 RPM
    - 9.3 watts (idle)
    - 8.5 ms avg. seek
    - 78 MB/s transfer rate
    - 5 year warrantee
    - $350 = $0.46 / GB

- Uses Perpendicular Magnetic Recording (PMR)!!
  - What’s that, you ask?

*source: www.seagate.com*
1 inch disk drive!

• Hitachi 2007 release
  • Development driven by iPods & digital cameras
  • 20GB, 5-10MB/s (higher?)
  • 42.8 x 36.4 x 5 mm

• Perpendicular Magnetic Recording (PMR)
  • FUNDAMENTAL new technique
  • Evolution from Logitudinal
    ▪ Starting to hit physical limit due to superparamagnetism
  • They say 10x improvement

www.hitachi.com/New/cnews/050405.html
www.hitachigst.com/hdd/research/recording_head/pr/
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 µdrives 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.

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

- iPod
- nano
- shuffle

Toshiba 1.8-inch HDD
30, 80GB

Samsung flash
2, 4, 8GB

Toshiba flash
1GB

en.wikipedia.org/wiki/Ipod
www.apple.com/ipod
## Upcoming Calendar

<table>
<thead>
<tr>
<th>Week #</th>
<th>Mon</th>
<th>Wed</th>
<th>Thu Lab</th>
<th>Fri</th>
</tr>
</thead>
<tbody>
<tr>
<td>#14</td>
<td>I/O Networks (Sameer)</td>
<td>I/O Disks</td>
<td>I/O Polling</td>
<td>Performance (Aaron)</td>
</tr>
<tr>
<td>Next week</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#15</td>
<td>Parallel Computing in Software</td>
<td>Parallel Computing in Hardware (Scott)</td>
<td>I/O Networking &amp; 61C Feedback Survey</td>
<td>LAST CLASS</td>
</tr>
<tr>
<td>Old Week</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#16</td>
<td>Sun 2pm Review 10 Evans</td>
<td></td>
<td>FINAL EXAM THU 12-14 @ 12:30pm-3:30pm 234 Hearst Gym</td>
<td></td>
</tr>
</tbody>
</table>

Garcia, Fall 2006 © UCB
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”
  - Low End → High End

Disk Array:
- 1 disk design
  - 3.5”
### 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>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. 9X</td>
</tr>
<tr>
<td>Power</td>
<td>3 KW</td>
<td>11 W</td>
<td>1 KW   3X</td>
</tr>
<tr>
<td>Data Rate</td>
<td>15 MB/s</td>
<td>1.5 MB/s</td>
<td>120 MB/s 8X</td>
</tr>
<tr>
<td>I/O Rate</td>
<td>600 I/Os/s</td>
<td>55 I/Os/s</td>
<td>3900 I/Os/s 6X</td>
</tr>
<tr>
<td>MTTF</td>
<td>250 KHrs</td>
<td>50 KHrs</td>
<td>??? Hrs</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?
Array Reliability

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

- Reliability of N disks
  \[ \text{Reliability of N disks} = \text{Reliability of 1 Disk} \div N \]
  (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

• 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-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 > 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
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 (RAID 2 has bit-level striping)

- 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-striping)

• 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
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
Peer Instruction Answer

1. **All** RAID (0-5) helps with performance, only RAID0 doesn’t help availability. TRUE

2. Surely! Must buy 2x disks rather than 1.25x (from diagram, in practice even even less) FALSE

3. RAID5 (2R,2W) vs. RAID1 (2W). Latency worse, throughput (I/O writes) better. TRUE

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 on seek, rotation improvements, MB/$ 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 $
  - Adds option for small # of extra disks
  - Can nest RAID levels
  - Today RAID is > tens-billion dollar industry, 80% nonPC disks sold in RAIDs, started at Cal
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.
  - 99.999% of the head/arm weight is supported by the air bearing force (air cushion) developed between the disk and the head.
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 jarring 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

NATIONAL GEOGRAPHIC • OCTOBER 2003

ART BY GRIFF WASON
Historical Perspective

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

- Form factor evolution:
  
  1970s: Mainframes $\Rightarrow$ 14 inch diameter disks
  
  1980s: Minicomputers, Servers $\Rightarrow$ 8”, 5.25” diameter disks
  
  Late 1980s/Early 1990s:
  
  - PCs $\Rightarrow$ 3.5 inch diameter disks
  - Laptops, notebooks $\Rightarrow$ 2.5 inch disks
  - Palmtops didn’t use disks, so 1.8 inch diameter disks didn’t make it
  
- Early 2000s:
  
  - MP3 players $\Rightarrow$ 1 inch disks
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”
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 / (100 MB/s) + 0.1 ms

= 8.5 ms + 0.5 /(7200 RPM/(60000ms/M))
  + 0.5 KB / (100 KB/ms) + 0.1 ms

= 8.5 + 4.17 + 0.005 + 0.1 ms = 12.77 ms

• How many CPU clock cycles is this?