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)

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(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. of 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, grown so fast that # of platters has reduced (some even use only 1 now!)
- 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: BPI (BPI)
  - # of tracks per surface: TPI
  - We care about bit density per unit area BPI x TPI
  - Called Areal Density = BPI x TPI
  - "~120 Gb/in² is longitudinal limit"
  - "230 Gb/in² now with perpendicular"
  - GB/$: > 100%/year (2X / 1.0 yrs)
    - Fewer chips + areal density

State of the Art: Two camps (2006)

- Performance
  - *Enterprise apps, servers*
    - E.g., Seagate Cheetah 15K.5
    - Ultra320 SCSI, 3 Gb/sec, Serial Attached SCSI (SAS)
    - 4Gb/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 warranty
    - $1000 = $3.30 / GB
- Capacity
  - Mainstream, home uses
    - E.g., Seagate Barracuda 7200.10
    - Serial ATA 3Gb/s (SATA/300), Ultra ATA 1/0
    - 750 GB, 2.5-inch disk
    - 7,200 RPM
    - 9.3 watts (idle)
    - 8.5 ms avg. seek
    - 78 MB/s transfer rate
    - 5 year warranty
    - $350 = $0.46 / GB
- Uses Perpendicular Magnetic Recording (PMR)!!
  - What's that, you ask?

Source: www.seagate.com

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.

Source: en.wikipedia.org/wiki/Flash_memory

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
  - source: www.hitachi.com/New/cnews/050405.html
  - www.hitachigst.com/hdd/research/recording_head/pr/
What does Apple put in its iPods?

- Samsung flash 2, 4, 8GB
- Toshiba 1.8-inch HDD 30, 80GB
- Toshiba flash 1GB

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

Disk Array: 1 disk design

Replace Small Number of Large Disks with Large Number of Small Disks! (1988 Disks)

<table>
<thead>
<tr>
<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>
</tr>
<tr>
<td>Volume</td>
<td>97 cu. ft.</td>
<td>0.1 cu. ft.</td>
</tr>
<tr>
<td>Power</td>
<td>3 KW</td>
<td>11 W</td>
</tr>
<tr>
<td>Data Rate</td>
<td>15 MB/s</td>
<td>1.5 MB/s</td>
</tr>
<tr>
<td>I/O Rate</td>
<td>600 I/Os/s</td>
<td>55 I/Os/s</td>
</tr>
<tr>
<td>MTTF</td>
<td>250 Khrs</td>
<td>50 Khrs</td>
</tr>
<tr>
<td>Cost</td>
<td>$250K</td>
<td>$2K</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
- 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

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

“And in conclusion…”

• Magnetic Disks continue rapid advance: 60%/yr capacity, 40%/yr bandwidth, slow on 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 $
  • 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/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.

Historical Perspective

- Form factor and capacity are more important in the marketplace than I/O 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

Early Disk History (IBM)

- Data density Mbit/sq. in. in Capacity of unit shown Megabytes
  1973: 1.7 Mbit/sq. in 140 MBytes
  1979: 7.7 Mbit/sq. in 2,300 MBytes

Early Disk History

- 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?