Storage Devices and RAID

Professor David A. Patterson
Computer Science 252
Fall 2000

Outline

- Disk Basics
- Disk History
- Disk options in 2000
- Disk fallacies and performance
- Tapes
- RAID

---

Disk Basics

### Disk History

- **Cylinder**: all tracks under heads

### Disk options in 2000

- Tapes
- RAID

### Disk fallacies and performance

- 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**, 1/surface) over track ("seek"), select **surface**, wait for **sector** rotate under **head**, then read or write
  - "**Cylinder**": all tracks under heads

---

Photo of Disk Head, Arm, Actuator

- **Spindle**
- **Arm**
- **Head**
- **Actuator**
- **Platters (12)**
Disk Device Performance

- Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead
- Seek Time: depends on the number of tracks moved by the arm, and the seek speed of the disk
- Rotation Time: depends on the speed the disk rotates and how far the sector is from the head
- Transfer Time: depends on the data rate (bandwidth) of the disk (bit density) and the size of the request

Data Rate: Inner vs. Outer Tracks

- To keep things simple, originally kept the same number of sectors per track
  - Since outer tracks are longer, lower bits per inch
- Competition decided to keep BPI the same for all tracks ("constant bit density")
  - More capacity per disk
  - More sectors per track towards the edge
  - Since the disk spins at constant speed, outer tracks have faster data rate
- Bandwidth outer track 1.7X inner track!

Devices: Magnetic Disks

- Purpose:
  - Long-term, nonvolatile storage
  - Large, inexpensive, slow level in the storage hierarchy
- Characteristics:
  - Seek Time (~8 ms avg)
    - Positional latency
    - Rotational latency
  - Transfer rate
    - 10-30 MByte/sec
  - Blocks
- Capacity
  - Gigabytes
  - Quadruples every 3 years (aerodynamics)

Average distance sector from head?

- 1/2 time of a rotation
  - 7200 Revolutions Per Minute \(\Rightarrow\) 120 Rev/sec
  - 1 revolution = 1/120 sec \(\Rightarrow\) 8.33 milliseconds
  - 1/2 rotation (revolution) \(\Rightarrow\) 4.16 ms

Average no. tracks move arm?

- Sum all possible seek distances from all possible tracks / # possible
  - Assumes average seek distance is random
  - Disk industry standard benchmark
### Historical Perspective

- **1956 IBM Ramac — early 1970s Winchester**
  - Developed for mainframe computers, proprietary interfaces
  - Steady shrink in form factor: 27 in. to 14 in.

- **1970s developments**
  - 5.25 inch floppy disk formfactor (microcode into mainframe)
  - Early emergence of industry standard disk interfaces
    - ST506, SASI, SMD, ESDI

- **Early 1980s**
  - PCs and first generation workstations

- **Mid 1980s**
  - Client/server computing
  - Centralized storage on file server
    - Accelerates disk downsizing: 8 inch to 5.25 inch
  - Mass market disk drives become a reality
    - Industry standards: SCSI, IPI, IDE
    - 5.25 inch drives for standalone PCs, End of proprietary interfaces

### Disk History

- **1973:**
  - 1.7 Mbit/sq. in
  - 140 MBytes

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

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

- **1997:**
  - 1,450 Mbit/sq. in
  - 2300 MBytes

- **1997:**
  - 3,090 Mbit/sq. in
  - 8,100 MBytes

Disk Performance Model / Trends

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

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

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

- **MB/$**
  - >100%/year (2X / <1.5 yrs)
  - Fewer chips + areal density

State of the Art: Ultrastar 72ZX

- 73.4 GB, 3.5 inch disk
- 2¢/MB
- 10,000 RPM;
  - 3 ms = 1/2 rotation
- 11 platters, 22 surfaces
- 15,110 cylinders
- 7 Gbit/sq. in. areal den
- 17 watts (idle)
- 0.1 ms controller time
- 5.3 ms avg. seek
- 50 to 29 MB/s (internal)

Latency = Queuing Time + Controller time + Seek Time + Rotation Time + Size / Bandwidth per access per byte

Source: www.ibm.com; www.pricewatch.com; 2/14/00

Disk Performance Example (will fix later)

- Calculate time to read 1 sector (512B) for UltraStar 72 using advertised performance; sector is on outer track

\[
\text{Disk latency} = \text{average seek time} + \text{average rotational delay} + \text{transfer time} + \text{controller overhead} \\
= 5.3 \text{ ms} + 0.5 \text{ ms} / (10000 \text{ RPM}) \text{ + 0.5 KB / (50 MB/s) + 0.15 ms} \\
= 5.3 \text{ ms} + 0.5 \text{ ms} / (10000 \text{ RPM} / (60000 \text{ ms/M})) \text{ + 0.5 KB / (50 KB/ms) + 0.15 ms} \\
= 5.3 + 3.0 + 0.10 + 0.15 \text{ ms} = 8.55 \text{ ms}
\]

Areal Density

- Bits recorded along a track
  - Metric is Bits Per Inch (BPI)
- Number of tracks per surface
  - Metric is Tracks Per Inch (TPI)
- Care about bit density per unit area
  - Metric is Bits Per Square Inch
  - Called Areal Density
  - Areal Density = BPI x TPI
Areal Density

- Areal Density = BPI x TPI
- Change slope 30%/yr to 60%/yr about 1991

Historical Perspective

- Form factor and capacity drives market, more than performance
- 1970s: Mainframes ⇒ 14 inch diameter disks
- 1980s: Minicomputers, Servers ⇒ 8”, 5.25” diameter disks
- Late 1980s/Early 1990s:
  - Pizzabox 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

### Disk Characteristics in 2000

<table>
<thead>
<tr>
<th>Disk parameter</th>
<th>Seagate Cheetah ST173404LC Ultra160 SCSI</th>
<th>IBM Travelstar 32GH DJSA - 232 ATA-4</th>
<th>IBM 1GB Microdrive DSCM-11000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk diameter (inches)</td>
<td>3.5</td>
<td>2.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Formatted data capacity (GB)</td>
<td>73.4</td>
<td>32.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Cylinders</td>
<td>14,100</td>
<td>21,664</td>
<td>7,167</td>
</tr>
<tr>
<td>Disks</td>
<td>12</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Recording Surfaces (Heads)</td>
<td>24</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Bytes per sector</td>
<td>512 to 4096</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>Avg Sectors per track (512 byte)</td>
<td>~ 424</td>
<td>~ 360</td>
<td>~ 140</td>
</tr>
<tr>
<td>Max. areal density (Gbit/sq.in.)</td>
<td>6.0</td>
<td>14.0</td>
<td>15.2</td>
</tr>
</tbody>
</table>

### Disk Characteristics in 2000

<table>
<thead>
<tr>
<th>Disk parameter</th>
<th>Seagate Cheetah ST173404LC Ultra160 SCSI</th>
<th>IBM Travelstar 32GH DJSA - 232 ATA-4</th>
<th>IBM 1GB Microdrive DSCM-11000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rotation speed (RPM)</td>
<td>10033</td>
<td>5411</td>
<td>3600</td>
</tr>
<tr>
<td>Avg. seek ms (read/write)</td>
<td>5.6/6.2</td>
<td>12.0</td>
<td>12.0</td>
</tr>
<tr>
<td>Minimum seek ms (read/write)</td>
<td>0.6/0.9</td>
<td>2.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Max. seek ms</td>
<td>14.0/15.0</td>
<td>23.0</td>
<td>19.0</td>
</tr>
<tr>
<td>Data transfer rate MB/second</td>
<td>27 to 40</td>
<td>11 to 21</td>
<td>2.6 to 4.2</td>
</tr>
</tbody>
</table>

### Disk Characteristics in 2000

<table>
<thead>
<tr>
<th>Disk parameter</th>
<th>Seagate Cheetah ST173404LC Ultra160 SCSI</th>
<th>IBM Travelstar 32GH DJSA - 232 ATA-4</th>
<th>IBM 1GB Microdrive DSCM-11000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buffer size in MB</td>
<td>4.0</td>
<td>2.0</td>
<td>0.125</td>
</tr>
<tr>
<td>Size: height x width x depth</td>
<td>1.6 x 4.0 x 5.8</td>
<td>0.5 x 2.7 x 3.9</td>
<td>0.2 x 1.4 x 1.7</td>
</tr>
<tr>
<td>Weight pounds</td>
<td>2.00</td>
<td>0.34</td>
<td>0.035</td>
</tr>
<tr>
<td>Rated MTTF in powered-on hours</td>
<td>1,200,000</td>
<td>(300,000?)</td>
<td>(20K/5 yr life?)</td>
</tr>
<tr>
<td>% of POH per month</td>
<td>100%</td>
<td>45%</td>
<td>20%</td>
</tr>
<tr>
<td>% of POH seeking, reading, writing</td>
<td>90%</td>
<td>20%</td>
<td>20%</td>
</tr>
</tbody>
</table>

### Disk Characteristics in 2000

<table>
<thead>
<tr>
<th>Disk parameter</th>
<th>Seagate Cheetah ST173404LC Ultra160 SCSI</th>
<th>IBM Travelstar 32GH DJSA - 232 ATA-4</th>
<th>IBM 1GB Microdrive DSCM-11000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load/Unload cycles (disk powered on/off)</td>
<td>&lt;1 per 10^{15}</td>
<td>&lt; 1 per 10^{13}</td>
<td>&lt; 1 per 10^{13}</td>
</tr>
<tr>
<td>Nonrecoverable read errors per bits read</td>
<td>not available</td>
<td>not available</td>
<td>not available</td>
</tr>
<tr>
<td>Seek errors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shock tolerance: Operating, Not operating</td>
<td>10 G, 175 G</td>
<td>150 G, 700 G</td>
<td>175 G, 1500 G</td>
</tr>
<tr>
<td>Vibration tolerance: Operating, Not operating (sine swept, 0 to peak)</td>
<td>5-400 Hz @ 0.5G, 22-400 Hz @ 1.0G, 2.5-500 Hz @ 2.0G</td>
<td>5-500 Hz @ 1G, 1-500 Hz @ 2.0G, 2.5-500 Hz @ 2.0G</td>
<td>500 Hz @ 5G</td>
</tr>
</tbody>
</table>
**Technology Trends**

- Today: Processing Power Doubles Every 18 months
- Today: Memory Size Doubles Every 18-24 months (4X/3yr)
- Today: Disk Capacity Doubles Every 12-18 months
- Disk Positioning Rate (Seek + Rotate) Doubles Every Ten Years!

---

**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 laid out on disk, where seek in real applications, then measure performance
  - 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)
  - UltraStar 72 avg. seek: 5.3 ms ⇒ 1.7 ms

---

**Fallacy: Use Data Sheet Transfer Rate**

- Manufacturers quote the speed off 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 3/4 of internal media rate (1.3X slower) for data
- For example, UltraStar 72 quotes 50 to 29 MB/s internal media rate
  ⇒ Expect 37 to 22 MB/s user data rate

---

**Disk Performance Example**

- Calculate time to read 1 sector for UltraStar 72 again, this time using 1/3 quoted seek time, 3/4 of internal outer track bandwidth; (8.55 ms before)

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

= (0.33 * 5.3 ms) + 0.5 * 1/(10000 RPM) + 0.5 KB / (0.75 * 50 MB/s) + 0.15 ms

= 1.77 ms + 0.5 / (10000 RPM/(60000ms/M)) + 0.5 KB / (37 KB/ms) + 0.15 ms

= 1.73 + 3.0 + 0.14 + 0.15 ms = 5.02 ms
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
  - Year  Sequentially  Randomly (1 sector/seek)
  - 1990  4 minutes  6 hours
  - 2000  12 minutes  1 week(!)
- 3.5” form factor make sense in 5-7 yrs?

SCSI: Small Computer System Interface

- Clock rate: 5 MHz / 10 (fast) / 20 (ultra)- 80 MHz (Ultra3)
- Width: n = 8 bits / 16 bits (wide); up to n – 1 devices to communicate on a bus or “string”
- Devices can be slave (“target”) or master(“initiator”)
- SCSI protocol: a series of “phases”, during which specific actions are taken by the controller and the SCSI disks
  - Bus Free: No device is currently accessing the bus
  - Arbitration: When the SCSI bus goes free, multiple devices may request (arbitrate for) the bus; fixed priority by address
  - Selection: informs the target that it will participate (Reselection if disconnected)
  - Command: the initiator reads the SCSI command bytes from host memory and sends them to the target
  - Data Transfer: data in or out, initiator: target
  - Message Phase: message in or out, initiator: target (identify, save/restore data pointer, disconnect, command complete)
  - Status Phase: target, just before command complete

Tape vs. Disk

- Longitudinal tape uses same technology as hard disk; tracks its density improvements
- Disk head flies above surface, tape head lies on surface
- Disk fixed, tape removable

Inherent cost-performance based on geometries:
  - fixed rotating platters with gaps (random access, limited area, 1 media / reader)
  - removable long strips wound on spool (sequential access, “unlimited” length, multiple / reader)

New technology trend:
  - Helical Scan (VCR, Camcoder, DAT)
  - Spins head at angle to tape to improve density

Tape wear out:
  - Helical 100s of passes to 1000s for longitudinal

Head wear out:
  - 2000 hours for helical

Both must be accounted for in economic / reliability model

Long rewind, eject, load, spin-up times; not inherent, just no need in marketplace (so far)

Designed for archival
Automated Cartridge System

STC 4400

8 feet

10 feet

6000 x 30 GB D3 tapes = 180 TBytes in 2000
Library of Congress: all information in the world; in 1992, ASCII of all books = 30 TB

Library vs. Storage

- Getting books today as quaint as the way I learned to program
  - punch cards, batch processing
  - wander thru shelves, anticipatory purchasing
- Cost $1 per book to check out
- $30 for a catalogue entry
- 30% of all books never checked out
- Write only journals?
- Digital library can transform campuses
- Will have lecture on getting electronic information

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”

Advantages of Small Formfactor Disk Drives

Low cost/MB
High MB/volume
High MB/watt
Low cost/Actuator

Cost and Environmental Efficiencies
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 x70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>20 GBytes 320 MBytes 23 GBytes</td>
</tr>
<tr>
<td>Volume</td>
<td>97 cu. ft. 0.1 cu. ft. 11 cu. ft.</td>
</tr>
<tr>
<td>Power</td>
<td>3 KW 11 W 1 KW 3X</td>
</tr>
<tr>
<td>Data Rate</td>
<td>15 MB/s 1.5 MB/s 120 MB/s 8X</td>
</tr>
<tr>
<td>I/O Rate</td>
<td>600 I/Os/s 55 I/Os/s 3900 I/Os/s 6X</td>
</tr>
<tr>
<td>MTTF</td>
<td>250 Khrs 50 Khrs ??? Khrs</td>
</tr>
<tr>
<td>Cost</td>
<td>$250K $2K $150K</td>
</tr>
</tbody>
</table>

Array Reliability

- Reliability of N disks = Reliability of 1 Disk ÷ N
  50,000 Hours ÷ 70 disks = 700 hours
- Disk system MTTF: Drops from 6 years to 1 month!
- Arrays (without redundancy) too unreliable to be useful!
- Hot spares support reconstruction in parallel with access: very high media availability can be achieved

Disk Arrays have potential for large data and I/O rates, 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

Redundant Arrays of Inexpensive Disks

RAID 1: Disk Mirroring/Shadowing

- Each disk is fully duplicated onto its “mirror”
- Very high availability can be achieved
- Bandwidth sacrifice on write:
  Logical write = two physical writes
- Reads may be optimized
- Most expensive solution: 100% capacity overhead
- (RAID 2 not interesting, so skip)
Redundant Array of Inexpensive Disks (RAID)

### RAID 3
- Sum computed across recovery group to protect against hard disk failures, stored in P disk.
- Logically, a single high capacity, high transfer rate disk: good for large transfers.
- Wider arrays reduce capacity costs, but decreases availability.
- 33% capacity cost for parity in this configuration.

#### RAID 3 Example
- Logical record: 1 1 1 1
- Striped physical records: 1 0 1 0
- P contains sum of other disks per stripe: 0 0 0 0
- If disk fails, subtract P from sum of other disks to find missing information.

---

#### RAID 4 Inspiration
- RAID 3 relies on parity disk to discover errors on read.
- But every sector has an error detection field.
- Rely on error detection field to catch errors on read, not on the parity disk.
- Allows independent reads to different disks simultaneously.

---

#### Redundant Arrays of Inexpensive Disks (RAID 4)
- High I/O Rate Parity
- Example: small read D0 & D5, large write D12-D15.
Inspiration for RAID 5
- RAID 4 works well for small reads
- Small writes (write to one disk):
  - Option 1: read other data disks, create new sum and write to Parity Disk
  - Option 2: since P has old sum, compare old data to new data, add the difference to P
- Small writes are limited by Parity Disk: Write to D0, D5 both also write to P disk

RAID 4 works well for small reads.

Independent writes possible because of interleaved parity.

Example: write to D0, D5 uses disks 0, 1, 3, 4

RAID-5: Small Write Algorithm
1 Logical Write = 2 Physical Reads + 2 Physical Writes

System Availability: Orthogonal RAIDs

Data Recovery Group: unit of data redundancy
Redundant Support Components: fans, power supplies, controller, cables
End to End Data Integrity: internal parity protected data paths
System-Level Availability

Fully dual redundant

I/O Controller

Array Controller

I/O Controller

Array Controller

with duplicated paths, higher performance can be obtained when there are no failures

Recovery Group

Goal: No Single Points of Failure

Summary: Redundant Arrays of Disks (RAID) Techniques

- **Disk Mirroring, Shadowing (RAID 1)**
  - Each disk is fully duplicated onto its "shadow"
  - Logical write = two physical writes
  - 100% capacity overhead

- **Parity Data Bandwidth Array (RAID 3)**
  - Parity computed horizontally
  - Logically a single high data bw disk

- **High I/O Rate Parity Array (RAID 5)**
  - Interleaved parity blocks
  - Independent reads and writes
  - Logical write = 2 reads + 2 writes
  - Parity + Reed-Solomon codes

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 $19 billion dollar industry, 80% nonPC disks sold in RAIDs