Storage

Jin-Soo Kim (jinsookim@skku.edu)
Computer Systems Laboratory
Sungkyunkwan University
http://csl.skku.edu
Hard Disk (1)

- Nonvolatile, rotating magnetic storage
## Hard Disk (2)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Seagate ST33000655SS</th>
<th>Seagate ST31000340NS</th>
<th>Seagate ST973451SS</th>
<th>Seagate ST9160821AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk diameter (inches)</td>
<td>3.50</td>
<td>3.50</td>
<td>2.50</td>
<td>2.50</td>
</tr>
<tr>
<td>Formatted data capacity (GB)</td>
<td>147</td>
<td>1000</td>
<td>73</td>
<td>160</td>
</tr>
<tr>
<td>Number of disk surfaces (heads)</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Rotation speed (RPM)</td>
<td>15,000</td>
<td>7200</td>
<td>15,000</td>
<td>5400</td>
</tr>
<tr>
<td>Internal disk cache size (MB)</td>
<td>16</td>
<td>32</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>External interface, bandwidth (MB/sec)</td>
<td>SAS, 375</td>
<td>SATA, 375</td>
<td>SAS, 375</td>
<td>SATA, 150</td>
</tr>
<tr>
<td>Sustained transfer rate (MB/sec)</td>
<td>73–125</td>
<td>105</td>
<td>79–112</td>
<td>44</td>
</tr>
<tr>
<td>Minimum seek (read/write) (ms)</td>
<td>0.2/0.4</td>
<td>0.8/1.0</td>
<td>0.2/0.4</td>
<td>1.5/2.0</td>
</tr>
<tr>
<td>Average seek read/write (ms)</td>
<td>3.5/4.0</td>
<td>8.5/9.5</td>
<td>2.9/3.3</td>
<td>12.5/13.0</td>
</tr>
<tr>
<td>Mean time to failure (MTTF) (hours)</td>
<td>1,400,000 @ 25°C</td>
<td>1,200,000 @ 25°C</td>
<td>1,600,000 @ 25°C</td>
<td>—</td>
</tr>
<tr>
<td>Annual failure rate (AFR) (percent)</td>
<td>0.62%</td>
<td>0.73%</td>
<td>0.55%</td>
<td>—</td>
</tr>
<tr>
<td>Contact start-stop cycles</td>
<td>—</td>
<td>50,000</td>
<td>—</td>
<td>&gt;600,000</td>
</tr>
<tr>
<td>Warranty (years)</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Nonrecoverable read errors per bits read</td>
<td>&lt;1 sector per 10^{16}</td>
<td>&lt;1 sector per 10^{16}</td>
<td>&lt;1 sector per 10^{16}</td>
<td>&lt;1 sector per 10^{14}</td>
</tr>
<tr>
<td>Temperature, shock (operating)</td>
<td>5°–55°C, 60 G</td>
<td>5°–55°C, 63 G</td>
<td>5°–55°C, 60 G</td>
<td>0°–60°C, 350 G</td>
</tr>
<tr>
<td>Size: dimensions (in.), weight (pounds)</td>
<td>1.0” × 4.0” × 5.8”, 1.5 lbs</td>
<td>1.0” × 4.0” × 5.8”, 1.4 lbs</td>
<td>0.6” × 2.8” × 3.9”, 0.5 lbs</td>
<td>0.4” × 2.8” × 3.9”, 0.2 lbs</td>
</tr>
<tr>
<td>Power: operating/idle/standby (watts)</td>
<td>15/11/—</td>
<td>11/8/1</td>
<td>8/5.8/—</td>
<td>1.9/0.6/0.2</td>
</tr>
<tr>
<td>GB/cu. in., GB/watt</td>
<td>6 GB/cu.in., 10 GB/W</td>
<td>43 GB/cu.in., 91 GB/W</td>
<td>11 GB/cu.in., 9 GB/W</td>
<td>37 GB/cu.in., 84 GB/W</td>
</tr>
<tr>
<td>Price in 2008, $/GB</td>
<td>~ $250, ~ $1.70/GB</td>
<td>~ $275, ~ $0.30/GB</td>
<td>~ $350, ~ $5.00/GB</td>
<td>~ $100, ~ $0.60/GB</td>
</tr>
</tbody>
</table>
Hard Disk (3)

- Hard disk internals

- Our Boeing 747 will fly at the altitude of only a few mm at the speed of approximately 65mph periodically landing and taking off.
- And still the surface of the runway, which consists of a few mm-thick layers, will stay intact for years.
Disk Sectors and Access

- **Each sector records**
  - Sector ID
  - Data (512 bytes, 4096 bytes proposed)
  - Error correcting code for defects & recording errors
  - Synchronization fields and gaps

- **Access to a sector involves**
  - Queuing delay if other accesses are pending
  - Seek: move the heads
  - Rotational delay
  - Data transfer
  - Controller overhead
**Disk Access Example**

- **Given**
  - 512B sector, 15,000rpm, 4ms average seek time, 100MB/s transfer rate, 0.2ms controller overhead, idle disk

- **Average read time**
  - 4ms seek time
    + \( \frac{1}{2} / (15,000/60) \) = 2ms rotational latency
    + \( 512 / 100\text{MB/s} \) = 0.005ms transfer time
    + 0.2ms controller delay
    = 6.2ms

- **If actual average seek time is 1ms**
  - Average read time = 3.2ms
Disk Performance Issues

- Manufacturers quote average seek time
  - Based on all possible seeks
  - Locality and OS scheduling lead to smaller actual average seek times

- Smart disk controller allocate physical sectors on disk
  - Present logical sector interface to host
  - SCSI, ATA, SATA

- Disk drives include caches
  - Prefetch sectors in anticipation of access
  - Avoid seek and rotational delay
RAID

- **Redundant Array of Inexpensive Disks**
  - Use multiple smaller disks as a cost-effective alternative to large expensive disks ($I = \text{Inexpensive}$)
  - Provide higher reliability and higher data transfer ($I = \text{Independent}$)
    - Parallelism improves performance
    - Plus extra disk(s) for redundant data storage
  - Improving performance via parallelism
    - Data striping: bit-level vs. block-level
  - Improving reliability via redundancy
    - Mirroring (shadowing)
    - Parity or error-correcting codes (ECCs)
RAID 0

- Non-redundant striping ("AID"?)
  - Data is broken into blocks
  - Each block is striped across multiple disks
  - I/O performance is greatly improved by spreading the I/O load across many channels and drives
  - Typically used in data rate intensive applications
    - e.g., video editing
RAID 1

- **Mirrored disks**
  - N + N disks
  - Expensive, highest disk overhead
  - Twice the read transaction rate of single disks, same write transaction rate as single disks
  - No rebuild is necessary in case of a disk failure
RAID 5

- Block-interleaved distributed parity
  - \( N + 1 \) disks
  - Like RAID 4, but parity blocks distributed across disks
  - Speed up small writes in multiprocessing systems, as the parity disk does not become a bottleneck
  - Widely used
Flash Memory

Jin-Soo Kim (jinsookim@skku.edu)
Computer Systems Laboratory
Sungkyunkwan University
http://csl.skku.edu
Flash Memory Characteristics

- **Erase-before-write**
  - Read
  - Write or Program – change state from 1 to 0
  - Erase – change state from 0 to 1

- **Bulk erase**
  - Program unit:
    - NOR: byte or word
    - NAND: sector or page
  - Erase unit: block
NOR vs. NAND Flash

- **NOR**
  - Random, direct access interface
  - Fast random reads
  - Slow erase and write
  - Mainly for code storage

- **NAND**
  - I/O mapped access
  - Smaller cell size
  - Lower cost
  - Better performance for erase and write
  - Mainly for data storage
NAND Technology (1)

- Density growth

![Density Growth Graph]

Source: Samsung Electronics

**Key Points**

- NAND Technology Driver: ~4x / 3 Years
- DRAM vs. NOR vs. SRAM growth patterns

---

ICE3003: Computer Architecture  | Spring 2012 | Jin-Soo Kim (jinsookim@skku.edu)
NAND Technology (2)

- Cost trends

![Graph showing cost trends for HDD, DRAM, and NAND technologies over time.](image)

Source: IEEE Computer, 2011
NAND Flash Architecture

- 2Gb NAND flash device organization

Source: Micron Technology, Inc.
NAND Flash Types (1)

- **SLC NAND flash**
  - Small block (≤ 1Gb)
  - Large block (≥ 1Gb)

- **MLC NAND flash**
  - 2 bits/cell

- **TLC NAND flash**
  - 3 bits/cell

Source: Micron Technology, Inc.
## NAND Flash Types (2)

<table>
<thead>
<tr>
<th></th>
<th>SLC NAND(^1) (small block)</th>
<th>SLC NAND(^2) (large block)</th>
<th>MLC NAND(^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page size (Bytes)</td>
<td>512+16</td>
<td>2,048+64</td>
<td>4,096+128</td>
</tr>
<tr>
<td>Pages / Block</td>
<td>32</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>Block size</td>
<td>16KB</td>
<td>128KB</td>
<td>512KB</td>
</tr>
<tr>
<td>(t_R) (read)</td>
<td>15 µs (max)</td>
<td>20 µs (max)</td>
<td>50 µs (max)</td>
</tr>
<tr>
<td>(t_{PROG}) (program)</td>
<td>200 µs (typ)</td>
<td>200 µs (typ)</td>
<td>600 µs (typ)</td>
</tr>
<tr>
<td></td>
<td>500 µs (max)</td>
<td>700 µs (max)</td>
<td>1,200 µs (max)</td>
</tr>
<tr>
<td>(t_{BERS}) (erase)</td>
<td>2 ms (typ)</td>
<td>1.5 ms (typ)</td>
<td>3 ms (typ)</td>
</tr>
<tr>
<td></td>
<td>3 ms (max)</td>
<td>2 ms (max)</td>
<td></td>
</tr>
<tr>
<td>NOP</td>
<td>1 (main), 2 (spare)</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Endurance Cycles</td>
<td>100K</td>
<td>100K</td>
<td>10K</td>
</tr>
<tr>
<td>ECC (per 512Bytes)</td>
<td>1 bit ECC 2 bits EDC</td>
<td>1 bit ECC 2 bits EDC</td>
<td>4 bits ECC 5 bits EDC</td>
</tr>
</tbody>
</table>

\(^1\) Samsung K9F1208X0C (512Mb)  \(^2\) Samsung K9K8G08U0A (8Gb)  \(^3\) Micron Technology Inc.
HDDs vs. SSDs (1)

2.5" HDD            Flash SSD
(101x70x9.3mm) 

1.8" HDD      Flash SSD
(78.5x54x4.15mm)

Top
 Bottom
# HDDs vs. SSDs (2)

<table>
<thead>
<tr>
<th>Feature</th>
<th>SSD (Samsung)</th>
<th>HDD (Seagate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>MMDOE56G5MXP (PM800)</td>
<td>ST9500420AS (Momentus 7200.4)</td>
</tr>
<tr>
<td>Capacity</td>
<td>256GB (16Gb MLC x 128, 8 channels)</td>
<td>500GB (2 Discs, 4 Heads, 7200RPM)</td>
</tr>
<tr>
<td>Form factor</td>
<td>2.5” Weight: 84g</td>
<td>2.5” Weight: 110g</td>
</tr>
<tr>
<td>Host interface</td>
<td>Serial ATA-2 (3.0 Gbps) Host transfer rate: 300MB</td>
<td>Serial ATA-2 (3.0 Gbps) Host transfer rate: 300MB</td>
</tr>
<tr>
<td>Power consumption</td>
<td>Active: 0.26W Idle/Standby/Sleep: 0.15W</td>
<td>Active: 2.1W (Read), 2.2W (Write) Idle: 0.69W, Standby/Sleep: 0.2W</td>
</tr>
<tr>
<td>Performance</td>
<td>Sequential read: Up to 220 MB/s Sequential write: Up to 185 MB/s</td>
<td>Power-on to ready: 4.5 sec Average latency: 4.17 msec</td>
</tr>
<tr>
<td>Measured performance¹</td>
<td>Sequential read: 176.73 MB/s Sequential write: 159.98 MB/s Random read: 10.56 MB/s Random write: 2.93 MB/s</td>
<td>Sequential read: 86.07 MB/s Sequential write: 84.64 MB/s Random read: 0.61 MB/s Random write: 1.28 MB/s</td>
</tr>
<tr>
<td>Price²</td>
<td>539,190 won</td>
<td>80,400 won</td>
</tr>
</tbody>
</table>

SSD Internals

![SSD Internals Diagram]

Diagram shows the SSD's internal components and their connections:
- **SRAM Controller**
- **ROM Controller**
- **ARM CPU** with an APB bridge
- **Flash Controller**
- **BM (Buffer Manager)**
- **SATA Device**
- **DRAM Controller**
- **Mem Utility**
- **NAND FLASH**
- **SATA Host**
- **DRAM**

Connections include:
- System Bus
- DRAM Access Bus

Components linked to system interfaces:
- **Clock Generator**
- **UART**
- **GPIO**
- **Timer**
- **WDC**
- **PMU**
- **ICU**
Flash Translation Layer (FTL)

- A software layer to make NAND flash fully emulate traditional block devices (or disks)

Source: Zeen Info. Tech.
Example (1)

- \( W = \langle \{1\}, \{2\}, \{8\}, \{1\}, \{2\}, \{12, 13\}, \{9\} \rangle \)
  - write (\( \{1\}, A \))
Example (2)

- \( W = \langle \{1\}, \{2\}, \{8\}, \{1\}, \{2\}, \{12, 13\}, \{9\} \rangle \)

  - write \( \{1\}, A \)
  - write \( \{2\}, B \)
Example (3)

- $W = \langle \{1\}, \{2\}, \{8\}, \{1\}, \{2\}, \{12,13\}, \{9\} \rangle$
  - write (\{1\}, A)
  - write (\{2\}, B)
  - write (\{8\}, C)
Example (4)

- \( W = \langle \{1\}, \{2\}, \{8\}, \{1\}, \{2\}, \{12,13\}, \{9\} \rangle \)
  - write \( \{1\}, A \)
  - write \( \{2\}, B \)
  - write \( \{8\}, C \)
  - write \( \{1\}, D \)
Example (5)

- \( W = \langle \{1\}, \{2\}, \{8\}, \{1\}, \{2\}, \{12,13\}, \{9\} \rangle \)
  - write \( \{1\}, A \)
  - write \( \{2\}, B \)
  - write \( \{8\}, C \)
  - write \( \{1\}, D \)
  - write \( \{2\}, E \)
Example (6)

- \( W = \langle \{1\}, \{2\}, \{8\}, \{1\}, \{2\}, \{12,13\}, \{9\} \rangle \)
  - write \( \{1\}, A \)
  - write \( \{2\}, B \)
  - write \( \{8\}, C \)
  - write \( \{1\}, D \)
  - write \( \{2\}, E \)
  - write \( \{12,13\}, FG \)
Example (7)

- \( W = \langle \{1\}, \{2\}, \{8\}, \{1\}, \{2\}, \{12, 13\}, \{9\} \rangle \)
  - write (\{1\}, A)
  - write (\{2\}, B)
  - write (\{8\}, C)
  - write (\{1\}, D)
  - write (\{2\}, E)
  - write (\{12, 13\}, FG)
  - write (\{9\}, H)
Try AndroBench!

AndroBench: Storage Benchmarking Tool

Measure your storage performance

Sequential Read: 46.33 MB/s
Sequential Write: 7.94 MB/s
Random Read: 7.9 MB/s, 2032.29 IOPS(4K)
Random Write: 0.24 MB/s, 63.93 IOPS(4K)
SQLite Insert: 21.15 TPS, 14.18 sec
SQLite Update: 48.44 TPS, 6.19 sec