Disks

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

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

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Seagate ST3000655SSS</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 bit read</td>
<td>&lt;1 sector per 10(^{16})</td>
<td>&lt;1 sector per 10(^{15})</td>
<td>&lt;1 sector per 10(^{16})</td>
<td>&lt;1 sector per 10(^{14})</td>
</tr>
<tr>
<td>Size: dimensions (in.), weight (pounds)</td>
<td>1.0&quot; * 4.0&quot; * 5.8&quot;, 1.5 lbs</td>
<td>1.0&quot; * 4.0&quot; * 5.8&quot;, 1.4 lbs</td>
<td>0.6&quot; * 2.8&quot; * 3.9&quot;, 0.5 lbs</td>
<td>0.4&quot; * 2.8&quot; * 3.9&quot;, 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>
Disk Storage (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.
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
    + $\frac{1}{2} / (15,000/60) = 2$ms rotational latency
    + $512 / 100$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
**Fault:**
- Failure of a component
- May or may not lead to system failure

```
Service accomplishment
Service delivered as specified

Restoration

Failure

Service interruption
Deviation from specified service
```
Dependability measures

- Reliability: mean time to failure (MTTF)
- Service interruption: mean time to repair (MTTR)
- Mean time between failures
  - MTBF = MTTF + MTTR
- Availability = MTTF / (MTTF + MTTR)
- Improving availability
  - Increase MTTF: fault avoidance, fault tolerance, fault forecasting
  - Reduce MTTR: improved tools and processes for diagnosis and repair
Sources of failures

<table>
<thead>
<tr>
<th>Operator</th>
<th>Software</th>
<th>Hardware</th>
<th>System</th>
<th>Year data collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>42%</td>
<td>25%</td>
<td>18%</td>
<td>Datacenter (Tandem)</td>
<td>1985</td>
</tr>
<tr>
<td>15%</td>
<td>55%</td>
<td>14%</td>
<td>Datacenter (Tandem)</td>
<td>1989</td>
</tr>
<tr>
<td>18%</td>
<td>44%</td>
<td>39%</td>
<td>Datacenter (DEC VAX)</td>
<td>1985</td>
</tr>
<tr>
<td>50%</td>
<td>20%</td>
<td>30%</td>
<td>Datacenter (DEC VAX)</td>
<td>1993</td>
</tr>
<tr>
<td>50%</td>
<td>14%</td>
<td>19%</td>
<td>U.S. public telephone network</td>
<td>1996</td>
</tr>
<tr>
<td>54%</td>
<td>7%</td>
<td>30%</td>
<td>U.S. public telephone network</td>
<td>2000</td>
</tr>
<tr>
<td>60%</td>
<td>25%</td>
<td>15%</td>
<td>Internet services</td>
<td>2002</td>
</tr>
</tbody>
</table>
RAID

- **Redundant Array of Inexpensive Disks**
  - Use multiple smaller disks as a cost-effective alternative to large expensive disks (I = Inexpensive)
  - Provide higher reliability and higher data transfer (I = 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 2

- Memory-style error-correcting codes (ECC)
  - N + E disks
  - Split data at bit level across N disks
  - Generate E-bit ECC
  - Too complex, not used in practice
RAID 3

- Bit-interleaved parity
  - $N + 1$ disks
  - Stripe parity is generated on writes, recorded on the parity disk and checked on reads
  - Less storage overhead than RAID 2
  - Cannot service multiple requests simultaneously
  - Not widely used
RAID 4 (1)

- Block-interleaved parity
  - N + 1 disks
  - Data striped across N disks at block level, redundant disk stores parity for a group of blocks
  - Very good read performance (same as RAID 0)
  - Writes require parity data be updated each time
  - No support for multiple simultaneous writes
RAID 4 (2)

- Writes: RAID 3 vs. RAID 4
RAID 5 (1)

- 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
RAID 5 (2)

- RAID 4 vs. RAID 5
RAID 6

- P + Q redundancy
  - N + 2 disks
  - Block-level interleaving with two parity blocks (P, Q)
  - Continue to execute read and write requests in the presence of any two concurrent disk failures
RAID 0+1 (or RAID 01)

- Mirrored stripes
  - \( N \times 2 \) disks (\( N > 1 \))
  - Data striped across \( N \) disks, and then mirrored to another, equivalent stripe
  - A single drive failure will cause the whole array to become, in essence, a RAID 0 array
RAID 1+0 (or RAID 10)

- Striped mirrors
  - 2*N disks (N > 1)
  - Data are mirrored in pairs, and then the resulting mirror pairs are striped
  - More reliable than RAID 0+1 → fault tolerant as long as no two disks are part of the same mirror
RAID Summary

- RAID can improve performance and availability
  - High availability requires hot swapping
- Assumes independent disk failures
  - Too bad if the building burns down!
- See “Hard Disk Performance, Quality and Reliability”
Fallacies (1)

- **Disk dependability**
  - If a disk manufacturer quotes MTTF as 1,200,000hr (140yr)
    - A disk will work that long
  - Wrong: this is the mean time to failure
    - What is the distribution of failures?
    - What if you have 1000 disks?
      - How many will fail per year?

Annual Failure Rate (AFR) = \( \frac{1000 \text{ disks} \times 8760 \text{ hrs/disk}}{1200000 \text{ hrs/failure}} \) = 0.73%
Fallacies (2)

- Disk failure rates are as specified
  - Studies of failure rates in the field
    - Schroeder and Gibson: 2% to 4% vs. 0.6% to 0.8%
    - Pinheiro, et al.: 1.7% (first year) to 8.6% (third year) vs. 1.5%
  - Why?

- Always 1GB = $2^{30}$ Bytes
  - For bandwidth, use 1GB = $10^9$ B
  - For storage, use 1GB = $10^9$ B
  - A 1TB disk provides 1,000,000,000 bytes
  - 1GB = $2^{30}$ B = 1,073,741,824 bytes
Fallacies (3)

- **Disk scheduling**
  - Best to let the OS schedule disk accesses
    - But modern drives deal with logical block addresses
    - Map to physical track, cylinder, sector locations
    - Also, blocks are cached by the drive
  - OS is unaware of physical locations
    - Reordering can reduce performance
    - Depending on placement and caching