Storage, Networks and other Peripherals

Preetam Ghosh
preetam.ghosh@usm.edu
## Technology in Memory Hierarchy

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Registers</td>
<td>cache</td>
<td>Main Memory</td>
<td>Back-up Memory</td>
</tr>
<tr>
<td>Typical size</td>
<td>&lt; 1KB</td>
<td>&lt; 16MB</td>
<td>&lt; 16GB</td>
<td>&gt; 100GB</td>
</tr>
<tr>
<td>Technology</td>
<td>custom memory, CMOS</td>
<td>on-chip, off-chip CMOS, SRAM</td>
<td>CMOS, DRAM, RAMBUS, SDRAM, RDRAM</td>
<td>magnetic disk, RAID, Flash Memory, optical disks</td>
</tr>
<tr>
<td>Access Time (ns)</td>
<td>0.25-0.5</td>
<td>0.5 – 25</td>
<td>1000-5000</td>
<td>5,000,000</td>
</tr>
<tr>
<td>BW (MB/sec)</td>
<td>20K-100K</td>
<td>5K-10K</td>
<td>1K-5K</td>
<td>0.02K-0.15K</td>
</tr>
<tr>
<td>Managed by</td>
<td>compiler</td>
<td>Hardware</td>
<td>operating system</td>
<td>OS/operator</td>
</tr>
<tr>
<td>Backed by</td>
<td>cache</td>
<td>Main memory</td>
<td>Back-up memory</td>
<td>Tape or CD</td>
</tr>
</tbody>
</table>
DRAM

- **DRAM**: Dynamic Random Access Memory
- **Single Transistor** to store one bit.
  - Reading the bit can disturb the information.
  - To prevent loss of information, periodic *refreshment* of the transistors are required.
  - All the bits in a row are refreshed simultaneously.
  - Every DRAM cell must access every row within certain time window. Typically about 8 msec.
  - Memory Controller includes hardware logic to perform this function.
DRAM

- The number of PINS are too many for access and that increases cost. So address bus is *multiplexed*. This reduces the number of address PINs to half.
  - 64 MB: Arranged in *Rectangular Matrix*.
  - Row: 16,384: 14 bits
  - Column: 16,384: 14 bits
  - First *RAS* (Row Access Strobe) is sent
  - Then *CAS* (Column Access Strobe is sent)
  - *Refresh cycle time* is combined (RAS + CAS) time
  - Number of steps in a complete refresh is *(Memory-Capacity)*$^{1/2}$
Memory Array can be composed of lower modularity memory blocks
SRAM: Static RAM

Uses 6 transistors to store one bit. The information of the transistor is not disturbed when read.

SRAM can go to low power or standby mode for long periods. On standby mode it needs very little power to retain the information.

SRAM design is focused on speed, so address line is not multiplexed.

The access time and cycle time are same.

8-16 times faster than DRAM

8-16 times more expensive than DRAM
Magnetic Disk

Non volatile storage. In use since 1965.

Two Roles:
1. Long term nonvolatile memory for files when programs are not running
2. A level of the memory hierarchy below main memory used as backing store for virtual memory when programs are in execution mode.
Magnetic Disk

Disk: Collection of Platters [1..12]
Speed of rotation: 3600 – 15,000 rpm.
Platters: Metal or glass disk coated with magnetic recording materials on both sides.
Disk Diameters: 1 inch – 3.5 inch (more than 90% in 2.5 – 3.5 inch)
Track: Disk surface is divided into concentric circles called track. 5000 – 30,000 tracks on each surface.
Sector: Each track is divided into sectors. Track might have 100 – 500 sectors. A sector is the smallest unit that is read or written. Typical size 512 bytes

Areal density = [Tracks/Inch] on a disk surface. [Bits/Inch] on a track
# Characteristics of Magnetic Disk

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Seagate Cheetah</th>
<th>IBM Travelstar</th>
<th>IBM Microdrive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter (inches)</td>
<td>3.5</td>
<td>2.5</td>
<td>1.0</td>
</tr>
<tr>
<td>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>Heads</td>
<td>24</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Bytes/sector</td>
<td>512-4096</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>Avg sectors/track (512 bytes)</td>
<td>424</td>
<td>360 (256-469)</td>
<td>140</td>
</tr>
<tr>
<td>Gb/sqin (areal density)</td>
<td>6.0</td>
<td>14.0</td>
<td>15.2</td>
</tr>
<tr>
<td>RPM</td>
<td>10,033</td>
<td>5411</td>
<td>3600</td>
</tr>
<tr>
<td>Avg Seek (read/write) ms</td>
<td>5.6/6.2</td>
<td>12.0</td>
<td>12.0</td>
</tr>
<tr>
<td>Max Seek (Read/write)</td>
<td>14.0/15.0</td>
<td>23.0</td>
<td>19.0</td>
</tr>
<tr>
<td>Min Seek (Read/write)</td>
<td>0.6/0.9</td>
<td>2.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Link speed to disk buffer (MB/s)</td>
<td>160</td>
<td>67</td>
<td>13</td>
</tr>
<tr>
<td>Data Transfer rate (MB/s)</td>
<td>27-40</td>
<td>11-21</td>
<td>2.6-4.2</td>
</tr>
<tr>
<td>Power idle/working (W)</td>
<td>16.4/23.5</td>
<td>2.0/2.6</td>
<td>0.5/0.8</td>
</tr>
<tr>
<td>Weight (pounds)</td>
<td>2.00</td>
<td>0.34</td>
<td>0.035</td>
</tr>
<tr>
<td>MTTF (powered-on-hours)</td>
<td>1.2 M</td>
<td>20K</td>
<td>8.8 K</td>
</tr>
<tr>
<td>Size HxWxD</td>
<td>1.6x4.0x5.8</td>
<td>0.5x2.7x3.9</td>
<td>0.2x1.4x1.7</td>
</tr>
</tbody>
</table>
Magnetic Disk

<table>
<thead>
<tr>
<th>Sector no.</th>
<th>Gap</th>
<th>Information bits</th>
<th>Gap</th>
<th>Next Sector no.</th>
<th>Misc fields</th>
</tr>
</thead>
</table>

Fixed size: typically 512 bytes
Includes error correction code

**Recording in a sector**

All tracks have same number of sectors: variable density recording

*Constant bit density*: Outer tracks have more sectors than inner tracks.
The constant bit density is not really correct. The bit density variation between outer sector and inner sector exists. The outer sector records 1.7 times more bit, though it is 2.1 times longer.

A moveable arm containing *read/write head* is located over each surface.

Arms for all surfaces are connected together and move in conjunction. So all arms are on the same track location on all surfaces. The term cylinder is used to refer to all tracks under the arms at a given point of time.

Groups of bits are recorded using *run length limited code*, that ensures that there is minimum and maximum number of bits in a group that the reader must decipher before seeing *synchronization signal*. 
Magnetic Disk

*Seek time:* To read or write a sector, the disk controller gives a command to move the arm over the proper track. This operation is called seek. And the time to move the arm to the desired track is called seek time.

*Rotation latency or rotation delay.* The time for the requested sector to rotate under the head is called rotation latency or rotation delay. The average latency, is the half way around the disk.

Disk speed: 10,000 rpm.

Average Rotation Latency = \( \frac{0.5}{10,000} \text{ rpm} = \frac{0.5 \times 60}{10,000} \text{ sec} = 3.0 \text{ msec.} \)

*Transfer time:* The time to transfer the data block. Normally, data block is the bits in a sector. This time depends upon
- Block size
- Disk size
- Rotation speed
- Recording density
- Speed of electronics connecting the disk to computer

2001 Transfer Rate: 3 MB per second for 3600 rpm 1 inch disk, 65 Mb per second for 15,000 rpm, 3.5 inch disk
Seek time

Current position of Head

Distance to be moved [D]

New position of Head

Velocity of Head

Settle time

Time to move head by a distance D

\[ \text{Time}_{\text{seek}} = T_{\text{min}} + \left[ \frac{D}{D_{\text{avg}}} \right] \cdot \left[ T_{\text{avg}} - T_{\text{min}} \right] \]

\( T_{\text{min}} \): Minimum Seek Time
\( T_{\text{max}} \): Maximum Seek time
\( T_{\text{avg}} \): Average Seek Time
Average Disk Access Time

\[ T = T_{avg\_seek} + T_{rotation\_latency} + T_{data\_transfer} + T_{control} \]

\[ T_{rotation\_latency} = \frac{0.5}{RPM} \text{ sec} \]

\[ T_{data\_transfer} = \frac{\text{No\_of\_bytes\_in\_a\_sector}}{\text{Data\_transfer\_in\_bytes\_per\_sec}} \text{ sec} \]
Disk Reliability

**MTBF : Mean time between Failure**

**MTTR : Mean Time to Repair**

**Module Availability = MTBF / [MTBF + MTTR]**

**Example:**
- 10 disks each rated MTBF = 1 M hours
- 1 SCSI Controller, MTBF = 0.5 M hours
- 1 Power Supply, MTBF = 0.2 M hours
- 1 Fan, MTBF = 0.2 M hours
- 1 SCSI Cable, MTBF = 1 M hours

**Series System:**

![Diagram of series system]

System failure rate = \( \frac{10}{10^6} + \frac{2}{10^6} + \frac{5}{10^6} + \frac{5}{10^6} + \frac{1}{10^6} \)

\[ = \frac{23}{10^6} \]

*System MTBF = \( \frac{10^6}{23} \) = 43,500 hours*
RAID Cluster

1980: Small form factor Hard Disk developed for PC
Berkeley Group proposed **RAID** (redundant arrays of inexpensive disks).

**Mirroring** was in use by fault tolerant computer group for reliability (idea of **RAID1**).
Thinking Machine had arrays of 32 data disks and 7 check disks checking ECC for correction. (idea of **RAID2**)
IBM filed a patent for rotating priority. (idea of **RAID5**)

1987: First RAID paper from Patterson, Gibson and Katz
1990: EMC announced RAID1 product
? Micropolis offered **RAID3**
? Compaq offered **RAID4**
? IBM, Data General, NCR offered RAID5

*Today it is $27B Market*
## RAID Levels

<table>
<thead>
<tr>
<th>Raid Levels</th>
<th>Minimum # of disk fault survived</th>
<th>Example data disk</th>
<th>Corresponding check disks</th>
<th>Company product</th>
</tr>
</thead>
<tbody>
<tr>
<td>0  No redundant striped</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>widely used</td>
</tr>
<tr>
<td>1  Mirrored</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>EMC, IBM</td>
</tr>
<tr>
<td>2  Memory style ECC</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>Storage Concepts</td>
</tr>
<tr>
<td>3  Bit Interleaved Parity</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>Network Appliance</td>
</tr>
<tr>
<td>4  Block interleaved parity</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>Storage Concepts</td>
</tr>
<tr>
<td>5  Block Interleaved distributed parity</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>widely used</td>
</tr>
<tr>
<td>6.  P + Q Redundancy</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
RAID 0 and 1

**RAID 0**: Data is stripped across multiple disks. Many disks operate at the same time, and data appear to software as from a large disk.

**RAID1**: Mirroring or shadowing of data. Whenever data is written on a disk, it is mirrored on another redundant disk. Two ways of mirroring; ‘*stripped mirror*(1+0) or ‘*mirror striped*(0+1).

4 disks worth of data, using 8 disks. Two choices: (1) 4 pair of RAID 1 disks and data stripped across the 4 pairs called RAID 1+0, (2) One pair of RAID 0 with 4 disks, data is stripped on 4 disks and duplicated on the other pair called RAID 0+1
RAID 3 Bit Interleaved Parity

Use 1 in \( N \) reliability, where \( N \) is the number of disks in a reliability group. The read and writes to go to all the disks in the group plus one extra disk that holds check information in case there is failure.

**Parity** is the sum of all data in the disks taking modulo 2 and then store it on the redundant disk. In case of failure, the content of the failure disk is extracted by using the redundant disk and the content of \((N-1)\) disks to meet the parity.

**Data recovery is slow.**

Arms of the disk drives are **synchronized**. The waiting time depends on the **longest seek time** of the disks.
RAID 4: Block Interleaved Parity

RAID 4: similar to RAID 3, but data access mechanism is different.

**RAID3**

**RAID4**

Bottleneck on Parity Disk
Can read and write small data efficiently
RAID5: Distributed Block Interleaved Parity

**RAID 5**: Distributes the Parity bit on all the disks to remove the bottleneck of the parity disk.