05 Storage system and I/O

05.01-05.03 Magnetic devices – RAID

- Magnetic devices
- Dependability
- RAID configurations

Magnetic disks

- Main functionality:
  - Long-term non-volatile storage
  - Virtual memory support
- Organization
### Capacity

- **Size**
  - 1 to 12 platters of 1 to 3.5 inches each
  - 2 recording surfaces per platter
  - 5,000 to 30,000 tracks per surface (cylinders)
  - 100 to 500 sectors per track
  - 512 bytes per sector
- **Capacity**
  - $12 \times 2 \times 30,000 \times 500 \times 512 > 100$ Gbytes

### Performance

- **Basic performance metrics:**
  - Rotation frequency: 3,600 to 15,000 RPM
  - Rotation latency: 3 ms (0.5/10,000RPM*60)
  - Seek time: 5 to 12 ms
  - Actual seek time lower than average due to locality
  - Transfer time: 3 to 65 MB/s
- **Performance improvements:**
  - Read-ahead buffer: 0.125 to 4 MB
  - Proprietary disk caching
- **Communication bottleneck:**
  - I/O bus: 80 to 320 MB/s
Technology

- Density:
  \[ \text{ArealDensity} = \frac{\text{Tracks on a disk surface} \times \text{Bits on a track}}{\text{Inch} \times \text{Inch}} \]
  - Improvement rate:
    - 29% before 1988
    - 60% from 1988 to 1996 (as fast as DRAM density increase)
    - 100% since 1996
  - Today’s density: 20Gbits/inch²
  - Today’s linear density: 0.05µm/bit

- Speed:
  \[ \text{Speed} = \text{TrackLength} \times \text{RotationFrequency} \]
  - Today’s speed (indicative):
    - \( 1 \text{m} \times 10,000 \text{RPM} \times 60 = 60 \text{Km/h} \)

### Cost

<table>
<thead>
<tr>
<th>DRAM</th>
<th>Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency</td>
<td>1</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>56</td>
</tr>
<tr>
<td>Cost</td>
<td>100</td>
</tr>
</tbody>
</table>
### Dependability

![Diagram](image)

Reliability: probability of contiguous service accomplishment (MTTF)
Availability: probability of service accomplishment MTTF/(MTTF+MTTR)

### Reliability improvements

- Fault avoidance
  - Prevent fault occurrence by construction
- Fault tolerance
  - Use redundancy to guarantee service accomplishment despite the occurrence of (some) faults
  - Hot spares / Hot swapping techniques
- Error removal
  - Minimize the presence of latent errors
- Error forecasting
  - Evaluate the presence and consequences of errors
Reliability improvements

- RAID
  - Redundant Arrays of Independent Disks
  - Redundant Arrays of Inexpensive Disks

<table>
<thead>
<tr>
<th>RAID level</th>
<th>Tolerated faults</th>
<th>Example disks</th>
<th>Check disks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Nonredundant striped</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>Mirrored</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Memory-style ECC</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>Bit-interleaved parity</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>Block-interleaved parity</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>Block-interleaved distributed parity</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>P+Q redundancy</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

- Notice: each disk is assumed to be able to detect its own failures

RAID 0

- No redundancy
- Striped data
- The collection of disks appears as a unique large disk
- Performance of large accesses is increased since many disks may operate at once
- The reliability of N disks is approx. 1/N that of a single disk
RAID 1

• Mirroring or shadowing
• Data are also written in a mirror/shadow disk working as a hot spare capable of hot swapping
• RAID 0-1 (mirrored stripes)
  – 2 arrays of N striped disks each
• RAID 1-0 (striped mirrors)
  – Stripe data among N pairs of RAID1 disks

RAID 2

• Protection group of N disks
• Use bit-level striping
• Use a single disk for error correction
  • Disk N stores the parity (modulo-2 sum) of all other disks
  • On a failure, the data stored in the faulty disk can be restored by subtracting the content of all other disks from that of the protection disk
RAID 3

- Protection group of N disks
- Use byte-level striping
- Use a single disk for error correction
  - Disk N stores the parity (modulo-2 sum) of all other disks
  - On a failure, the data stored in the faulty disk can be restored by subtracting the content of all other disks from that of the protection disk
- Mirroring is a special case of RAID 3

- Every write goes to all disks!
RAID 4

- Improves RAID 3 by:
  - reducing the number of accesses required on writes
    - Check which bits change in the parity sum
    - Update only modified bits in the parity disk
  - using block-level striping
  - Centralized parity write is still a bottleneck

RAID 5

- Distributed block-interleaved parity
RAID 6

- P+Q redundancy
- A first parity disk (P) enables recovering from single errors
- A second parity disk (Q) enables recovering from a second (simultaneous) error
- 6 accesses on 3 disks required on each write