High-Performance Storage Systems

Storage Technology Drivers
- Driven by the prevailing computing paradigm
  - 1950s: migration from batch to on-line processing
  - 1990s: migration to ubiquitous computing
  - computers in phones, books, cars, video cameras, ...
  - nationwide fiber optical network with wireless tails
- Today: digital media everywhere
  - digital forms of voice, picture, and video
  - data from scientific computing such as earthquake simulation, high energy physical experiments, bioinformatics
  - data from personal storage, web server, peer-to-peer storage, grid storage
- Effects on storage industry:
  - Embedded storage
    - smaller, cheaper, more reliable, lower power
  - Data utilities
    - high capacity, hierarchically managed storage

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
    - 10-40 MByte/sec
  - Blocks
  - Capacity
    - gigabytes
    - quadruples every 2 years

Seagate Barracuda 180
- 181.6 GB, 3.5 inch disk
- 12 platters, 24 surfaces
- 24,247 cylinders
- 7,200 RPM; (4.2 ms avg latency)
- 7.4/8.2 ms avg seek (r/w)
- 64 to 35 MB/s (internal)
- 0.1 ms controller time
- 10.3 watts (idle)
Disk Performance Factors

Actual disk seek and rotation time depends on the current head position.

- **Seek time**: how far is the head to the track?
  - Disk industry standard: assume random position of the head, e.g., average 8ms seek time.
  - In practice: disk accesses have locality.

- **Rotation time**: how far is the head to sector?
  - Can safely assume 1/2 of rotation time (disk keeps rotating)
  - 10000 Revolutions Per Minute ⇒ 166.67 Rev/sec
  - 1 revolution = 1/166.67 sec ⇒ 6.00 ms
  - 1/2 rotation (revolution) ⇒ 3.00 ms

- **Data Transfer time**: What are the rotation speed, disk density, and sectors per transfer?
  - 10000 RPM ⇒ 1 track of data per 6.00 ms
  - Outer tracks are longer and may support higher bandwidth.

Disk Performance Example

- **Rule of Thumb**: Observed average seek time is typically about 1/4 to 1/3 of quoted seek time (i.e., 3X-4X faster).
- **Rule of Thumb**: Disks deliver about 3/4 of internal media rate (1.3X slower) for data.

- Calculate time to read 64 KB for UltraStar 72, using 1/3 quoted 7.4ms seek time, 3/4 of 64MB/s internal outer track bandwidth:

  \[
  \text{Disk latency} = \text{average seek time} + \text{average rotational delay} + \text{transfer time} + \text{controller overhead}
  \]

  \[
  = \left(0.33 \times 7.4 \text{ ms}\right) + 0.5 \times \frac{64 \text{ KB}}{65 \text{ MB/s}} + 1.0 \text{ ms}
  \]

  \[
  = 2.5 \text{ ms} + 0.5 \times 6.00 \text{ ms} + 1.0 \text{ ms}
  \]

  \[
  = 2.5 + 4.2 + 14.1 = 8.2 \text{ ms} (64\% of 12.7)
  \]

Disk Characteristics in 2000

<table>
<thead>
<tr>
<th></th>
<th>Seagate Cheetah 7200 ST373404LC</th>
<th>IBM Travelstar 320GB 3320 DJSA-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>
<tr>
<td></td>
<td>$828</td>
<td>$447</td>
<td>$435</td>
</tr>
</tbody>
</table>

Disk Performance/Cost Trends

- **Capacity**: + 100%/year (2X / 10 yrs)
- **Transfer rate (BW)**: + 40%/year (2X / 2 yrs)
- **Rotation + Seek time**: ~ 8%/ year (1/2 in 10 yrs)
- **MB/$**: > 100%/year (2X / 10 yrs)
  - Fewer chips + areal density

- **Seagate 120GB Internal Hard Drive ST3120026A, $150 at staple (list price)**
- **Maxtor 120GB 8MB Cache Hard Drive $59.84 after rebate at OfficeDepot**

Disk System Performance

- **System-level Metrics**:
  - **Response Time**
  - **Throughput**

  \[
  \text{Response Time} = \text{Queue} + \text{Controller} + \text{Device Service time (s)}
  \]

  \[
  \text{Throughput} = \frac{\text{Total BW}}{100%}
  \]

How About Queuing Time?

- Queuing time can be the most significant one in disk response time.
- More interested in long term, steady state than in startup ⇒ Arrivals = Departures.
- **Little’s Law**: Mean number tasks in system = arrival rate \times mean response time.
- Applies to any system in equilibrium, as long as nothing in black box is creating or destroying tasks.
A Little Queuing Theory: Notation

- Queuing models assume state of equilibrium: input rate = output rate
- Notation:
  - \( r \): average number of arriving customers/second
  - \( T_{ser} \): average time to service a customer (traditionally \( \mu = 1 / T_{ser} \))
  - \( T_{q} \): average time/customer in queue = \( T_{ser} \times u / (1 - u) \)
  - \( T_{sys} \): average time/customer in system = \( T_{q} + T_{ser} \)
  - \( L_{q} \): average length of queue = \( r \times T_{q} \)
  - \( L_{sys} \): average length of system = \( r \times T_{sys} \)
- Little's Law: \( \text{Length} = \text{rate} \times \text{Time} \)

A Little Queuing Theory: Example

- Processor sends 50 x 8KB disk I/Os per sec, requests & service exponentially distrib., avg. disk service = 12 ms
- On average, how is the disk utilized?
  - What is the average number of requests in the queue?
  - What is the average time spent in the queue?
  - What is the average response time for a disk request?

Notation:
- \( r \): average number of arriving customers/second = 50
- \( T_{ser} \): average time to service a customer = 12 ms
- \( u \): server utilization (0..1): \( u = r \times T_{ser} = 50 \times 0.012 = 0.60 \)
- \( T_{q} \): average time/customer in queue = \( 12 \times 0.60 / (1 - 0.60) = 10 \times 0.60 / 0.40 \approx 15 \) ms
- \( T_{sys} \): average time/customer in system: \( T_{q} + T_{ser} = 30 \) ms
- \( L_{q} \): average length of queue = \( 50 \times 0.018 = 0.9 \) requests in queue
- \( L_{sys} \): average # tasks in system = \( 50 \times 0.030 = 1.5 \)

Look into textbook when you need to work on I/O

Array Reliability

- Reliability of \( N \) disks = Reliability of 1 Disk \( \times \) N
  - 50,000 Hours + 70 disks = 700 hours
  - Disk system MTTF: Drops from 6 years to 1 month
  - (MTTF: Mean Time to Failure)
- Arrays (without redundancy) too unreliable to be useful!
- Solution: RAID -- Redundant Arrays of Inexpensive Disks

RAID: The Idea

- Example: small read D0 & D5, large write D12-D15
  - RAID-3 shown

RAID 4: High I/O Rate Parity

- Increasing Logical Disk Address
  - Stripe
  - Disk Columns
**RAID 5: High I/O Rate Interleaved Parity**

Independent writes possible because of interleaved parity.

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

Increasing Logical Disk Addresses

No disk hot spot!

**Future Storage Trends**

- **Disks:**
  - Extraordinary advance in capacity/drive, $/GB
  - Currently 17 Gbit/sq. inch; can continue past 100 Gbit/sq. inch?
  - Bandwidth, seek time not keeping up: 3.5 inch form factor makes sense? 2.5 inch form factor in near future? 1.0 inch form factor in long term?

- **Tapes**
  - Old technique, no investment in innovation
  - Are they already dead?
  - What is a tapeless backup system?

- **Other Storage**
  - CD/DVD
  - Compact Flash, USB key storage, MRAM