Tiering in Today’s Disk Storage Systems

Session 09444
John Ticic
John Baker
IntelliMagic Inc.
John’s OK!
Objectives

- Introduction
- Modern Storage Subsystem overview
- Hard Disk Drive overview
  - FC/SATA/SAS
  - SSD overview
  - HDD/SSD Service Times – utilization!
- Application Service Times
- I/O Profiles
  - R/W, Random Sequential
- Where will SSD’s help
- Roadblocks to success and Alternatives
Who is IntelliMagic?

- The Storage Performance Company.
  - Since 1991 software solutions to hardware vendors.
  - Since 2005 to some of the largest end-user sites (small too!)

- Deep industry expertise: founder is Dr. Gilbert Houtekamer, MVS I/O Subsystems author (w/ Dr. P. Artis)

- Solutions:

- Services:
  - 4 Day Class: z/OS Storage Performance & Architecture
  - Performance Diagnosis Study
  - Disk Subsystem Sizing & Configuration Study
  - Replication Bandwidth Analysis
  - Volume Migration Planning
About Me

- 4 years as Performance Specialist with IntelliMagic
- 15 years of mainframe experience at a large international bank

Responsibilities included:
- Far too much SAS
- “Bill”/WLM: pre and post Goal Mode
- Set CPU weights and virtual storage parms
- Online/batch tuning (1000+ online transactions/sec and 75000 batch jobs per day)
- DASD tuning (VSAM buffering, striping, tune sort parms, manage and place ‘loved’ data)
- Designed and implemented synchronous remote copy in production for all 13000 production volumes
- According to IBM this was the largest GDPS in the world at the time
- 100% availability of the Production Sysplex for over 10 years
Modern Storage Systems
Disk Subsystem Architecture

- All vendors agree:
  - Front-end Controllers are specialized processors to connect to hosts or other subsystems (copy services)
  - Back-end Controllers are specialized processors to connect to disks
  - A large cache memory is required to provide good performance for reads and writes
  - A high-speed interconnect is essential (bus or switch)

- Two copies
  - Battery back-up & two copies are essential for all I/O to avoid that data written is lost
  - Provided in all enterprise class equipment
Front-end Director

- Provides connectivity between disk subsystem and hosts
- Cards support ESCON, SCSI, FICON Fibre, SAS and/or iSCSI sometimes FICON and Fibre with one card
- Implementations differ greatly in maximum data handling capability, especially for FICON and Fibre
- Even though ports are rated as (e.g.) 4 Gbit/s, no implementation achieves this speed due to overhead.
Processors and Cache

- Different implementations use different approaches
- All use cache to store
  - Recently used tracks and records
  - Recently written records
  - Pre-loaded tracks for sequential read
  - Some form of track descriptor tables to facilitate write operations without a disk access
  - Async copy information
Device Adapters

- Connect HDDs to internal Disk Subsystem resources
- Manage RAID operations, sometimes using cache memory for RAID computations
- Configured in pairs to provide redundancy if one adapter fails
- HDD interfaces include various generations of SCSI, SSA, FC-AL, SATA and SSD
- FC-AL switched back-end are gradually being replaced by SAS back-ends
Disk Technology
Drive Types

HDD

Access in milliseconds

SSD Flash is derived of byte addressable EEPROM

Access in microseconds
Drive Protocols

Command sets commonly used:

- **CKD CCWs for zSeries mainframe**
  - Very elaborate command set
  - Designed around error detection and recovery
  - One command at a time per device address

- **ATA for low-cost PC applications**
  - Designed by Western Digital in 1986
  - One command at a time up through ATA-3
  - Write cache enabled but no battery back-up

- **SCSI for higher performance server applications**
  - Based on Shugart Associated System Interface (1979) (SASI, Apple II)
  - Well defined command set
  - Tagged Command Queuing
## Protocols and Connections

<table>
<thead>
<tr>
<th>ATA</th>
<th>SCSI</th>
<th>Wiring</th>
<th>Transfer Rate (MB/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial</td>
<td>SATA</td>
<td>SAS: Serial Attached SCSI</td>
<td>600**</td>
</tr>
<tr>
<td>Fibre Arbitrated Loop, Fibre</td>
<td>FATA</td>
<td>FC-AL, FC</td>
<td>800</td>
</tr>
<tr>
<td>Over TCP/IP</td>
<td>AoE (ATA over Ethernet)</td>
<td>iSCSI, FCoE</td>
<td>1000</td>
</tr>
<tr>
<td>‘SSA’</td>
<td>SSA</td>
<td>Copper (Twister pair)</td>
<td>160</td>
</tr>
</tbody>
</table>

Drive Performance Characteristics

<table>
<thead>
<tr>
<th>Feature</th>
<th>HDD</th>
<th>SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protocol: decode commands</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Seek time: position head</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Latency: wait for record to pass head</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Data transfer</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sequential pre-load, caching</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Optimize access</td>
<td>For speed</td>
<td>For wear</td>
</tr>
</tbody>
</table>
## Latency: Rotational Delay

### Table: RPM and Latency (ms)

<table>
<thead>
<tr>
<th></th>
<th>RPM per min</th>
<th>RPM per sec</th>
<th>Latency (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3390-3</td>
<td>4200</td>
<td>70</td>
<td>7.2</td>
</tr>
<tr>
<td>Older SATA</td>
<td>6000</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td>SATA</td>
<td>7200</td>
<td>120</td>
<td>4.1</td>
</tr>
<tr>
<td>Most Fibre drives</td>
<td>10,000</td>
<td>167</td>
<td>3</td>
</tr>
<tr>
<td>High end Fibre drives</td>
<td>15,000</td>
<td>250</td>
<td>2</td>
</tr>
<tr>
<td>Solid State Drive</td>
<td>n/a</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

- Average delay is half a rotation
## Disk Service Times

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Seek</th>
<th>Latency</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATA</td>
<td>1.0?</td>
<td>9</td>
<td>4.1</td>
</tr>
<tr>
<td>10k RPM Fibre</td>
<td>0.3?</td>
<td>4.7</td>
<td>3</td>
</tr>
<tr>
<td>15k RPM Fibre</td>
<td>0.2?</td>
<td>3.6</td>
<td>2</td>
</tr>
<tr>
<td>10k RPM SAS</td>
<td>0.2?</td>
<td>2.6</td>
<td>3</td>
</tr>
<tr>
<td>SSD</td>
<td>0.2?</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Protocol time**
  - Very small < 0.5 ms
- **Average seek, assuming fully used HDD**
  - Range 3.6 – 10 ms depending on technology
- **Latency**
  - Range 2 – 5 ms
- **Data transfer for 512 bytes**
  - Very small
- **Total service time for read**
  - From 0.2 to 15 ms
Application Service Times
What is the user experience?

- Total Response time = CPU + I/O + Wait + Network

- CPU
  - Not fast enough – buy a 196!
  - Too many instructions – chase application people

- Wait
  - WLM priority?
  - Overcommitted resources (see #1)

- Network – always a great place to blame 😊

- Let’s break down our I/O time...
I/O Response Components

- Response = IOSQ + Pending + Connect + Disconnect

- IOSQ
  - Wait for local device (UCB) busy

- Pending
  - Wait for channel, subsystem, or device in use by other LPAR

- Connect
  - Time required to transfer data and commands to disk subsystem plus protocol overhead.

- Disconnect
  - Wait for information to be retrieved from disk (read), written to device (write) or to a secondary controller (copy services), or internal CU delays.
Where Will SSD’s Help?

Response time components
for all data
I/O Profiles
Backend Load Depends on Workload Characteristics

- **Random read hits** have no impact on backend
- **Random read misses** must be resolved by accessing a physical disk
  - Synchronous; service time matters
- **Random Writes** are cache hits, but must be written to the physical disks
  - Largest write overhead
  - Asynchronous
- **Sequential reads** are 100% cache hits, but need to access the physical disks for 100%
  - Asynchronous
- **Sequential writes** are 100% cache hits, but must be written to the physical disks
  - Can usually be optimized
  - Asynchronous
Questions you need to Answer

- Read/Write Ratio
- Cache hit %
- Sequential %
- RAID type
- Business Importance
Hypothetical Scenario
I/O’s per Transaction

- Let’s say a typical transaction requires 100 I/O’s
- Let’s take the average I/O response time of 2 ms from our chart
- But – only about .5 ms of that is Disconnect time
What’s my Real Disconnect Time?

- RMF reports the average disconnect
- This does not mean that all I/O’s experienced disconnect
- The reality is that cache hits experience none (of significance)
- Disconnect time for misses can be calculated

What is the actual disconnect time for cache misses with an average disconnect of .5 ms and a hit ratio of 95%?

\[
\text{DISC}_m = \frac{\text{RMFDISC}}{\text{MISS RATIO}}
\]

\[
.5 \div .05 = 10 \text{ ms}
\]

What does this mean for the actual response times of our I/O’s?

95% of the I/O’s experienced no Disc. While 5% experienced 10 ms (no I/O’s experienced .5 ms!)
What if I was on SSD’s?

- Potentially reduce 10 ms to <1 ms!
  - For 5% of I/O’s

- 95% of I/O’s are getting 1.5 ms response

- 5% are getting 2 ms
  - How to identify the candidates?
The Road to SSD and Alternatives
SSD Roadblocks

- $ per GB
  - SSD vs FC/SAS vs SATA
  - Should improve with competition
  - MLC!

- SSD’s per DA... per DSS
  - Throughput limitations

- TB per DSS footprint
  - Floor space
  - Opposes desired consolidation

- Complex to implement efficiently
Selecting SSD Candidates

- **Loved ones**
  - May be cache friendly = minimal benefit

- **Auto tiering**
  - Based on activity; may not be important to business
  - Analysis window and reaction time?

- **SMF/RMF**
  - Difficult and time consuming

- **Software**
  - Hardware Vendor, IBM, IntelliMagic
Auto Tiering Options

- **EMC FAST**
  - Distributed systems: FAST for Virtual Pools (FAST VP) looks good
  - Very granular “chunk” size – 7.5 MB
  - Mainframe: Volume-level only
  - Three Tiers: Flash, FC (10K and 15K), SATA

- **HDS HDT**
  - Interesting “chunk” size of 42 MB
  - Virtualization – good or bad?
  - Mainframe soon

- **IBM EasyTier**
  - 1 GB chunk size. Standard IBM “Extent” for many years
  - 2 Tiers (2 of SSD, FC/SAS, SATA)
  - Mainframe today
MLC is coming!

- Original “Enterprise” SSD was only Single Level Cell (SLC)
- Can handle many more writes
- About 10x cost of Multi-Level Cell (MLC)

- IBM and Hitachi GST have certified MLC for enterprise use

Alternatives

- **Software Striping**
  - SMS striping
  - Very Granular (track/CI)
  - Span DSS’s (more channels = more throughput)

- **Hardware Striping**
  - Volume spanning RAID ranks
  - Chunk size may vary

- **Balance!**
  - Measure volume/rank activity
  - HDD response grows with disk utilization
  - ROT: stay under 50%
  - Use RMF or vendor tools
Conclusions

- Back end HDD response is only one component of overall response and represents a very small portion of total I/O

- SSD = $$$ (MLC? = $)

- Controllers are not ready for wide-spread use

- Proper implementation is complex

- What is your current back end response?

- Are your users unhappy about response?
Thank You

Questions?

John.Ticic@intellimagic.net
John.Baker@intellimagic.net