SAS SSDs – Building Blocks for High-Performance Storage

Ulrich Hansen
Director of Market Development
HGST, a Western Digital company
Agenda

SSD Market Context

Managing SSD Endurance

SSD Performance Considerations

SAS SSDs – Enterprise SSD Building Block
Drivers for High-Performance Storage

Several system and application drivers are increasing the demand for high-performance storage solutions going forward.

- Increase in randomness of IOs at the storage device level
- Increase in average throughput requirements
- Increase in latency and command completion time requirements

Storage devices utilizing non-volatile memories are uniquely positioned to close the ‘IO Gap’ and deliver these high-performance storage solutions.
Impacts of Cloud & Multi-Tenancy

"Traditional" Workloads

- Sequential throughput and IOPs important
- Utilization can be low
- Broad mix of HDD solutions are sufficient for performance demands

Cloud Server Workload

- High capacity HDDs & more VMs drive random IO (and fragmentation)
- IOPs critical
- 24x7 Operation
- SSD and Enterprise-Class HDDs preferred

Sequential & Random Accesses

More Highly Randomized Accesses
Evolution of Tiering

More Tiers, Not Less.
Rigid Tiers Disappearing

Tier 0
SSD

Tier 1
10K & 15K
SAS/FC HDD

Tier 2
7200 RPM SATA/SAS
HDD

Tier 3
Tape

15K SAS HDD
Resurgence

10K SAS HDD

7200 RPM SATA/SAS
HDD

Cold Storage HDD

Tape
...And Then the Lines Didn’t Cross…

Enterprise SSDs will continue to carry a significant $/GB premium over Enterprise HDDs – SSDs will be deployed where performance justified.

Pricing Dynamics
- Price/GB curves for Flash and HDD are very similar
- Creates opportunity for sub-tiering
- Implies that SSD and Performance HDD will be complimentary for the foreseeable future
Enterprise Market — Growth in SSD & Capacity HDD

TAM by Volume

'11-'16 CAGR
Total: 12%


Units (M)

SSD
38%

Performance HDD
-3%

Capacity HDD
22%

Source: HGST

2012 Storage Developer Conference.

© 2012 HGST, a Western Digital company
SSD Segment Fragmenting

**Technology**
- SLC, eMLC, MLC
- NAND Lithography

**Form Factor**
- 3.5”, 2.5”, 1.8”
- Add-In Card

**Capacity**
- 100GB – 2TB

**Performance**
- Random IOPS 30K – 600K

**Protocol**
- SATA, SAS, SOP
- ATA: SATA
- NVMe

**Interface**
- SATA, SAS, PCIe

**Endurance**
- Drive Writes / Day for 5 years: 50, 25, 10, 3,

**Increasing:**
- Competition
- Differentiation
- Complexity
- Fragmentation
Enterprise SSD – Market Outlook

SATA revenue growth is expected to level off in 2012 as SAS and PCIe are driving SSD growth in storage systems and servers, respectively.
The Next Gen – Removing Storage Bottlenecks

Bandwidth per Port / Lane (Full Duplex)

<table>
<thead>
<tr>
<th></th>
<th>MB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCIe 2.0</td>
<td>1000</td>
</tr>
<tr>
<td>6Gb/s SATA</td>
<td>1500</td>
</tr>
<tr>
<td>6Gb/s SAS</td>
<td>2000</td>
</tr>
<tr>
<td>PCIe 3.0</td>
<td>2500</td>
</tr>
<tr>
<td>12Gb/s SAS</td>
<td>3000</td>
</tr>
</tbody>
</table>

Bandwidth for Typical Drive Configurations (Full Duplex)

<table>
<thead>
<tr>
<th></th>
<th>MB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>6Gb/s SATA</td>
<td></td>
</tr>
<tr>
<td>6Gb/s SAS, Dual Port</td>
<td></td>
</tr>
<tr>
<td>PCIe 2.0, 4 Lanes</td>
<td></td>
</tr>
<tr>
<td>12Gb/s SAS, Dual Port</td>
<td></td>
</tr>
<tr>
<td>PCIe 3.0, 4 Lanes</td>
<td></td>
</tr>
</tbody>
</table>

Not Shown: 4 port MultiLink SAS or 8 Lane PCIe

With Dual Port SAS and Multi-lane PCIe the bottleneck is not Storage I/O bandwidth

© 2012 HGST, a Western Digital company
Agenda

SSD Market Context

Managing SSD Endurance

SSD Performance Considerations

SAS SSDs – Enterprise SSD Building Block
Write Endurance – Why it Exists …

• Programming and erasing of a typical NAND cell involves forcing charge carriers through the dielectric isolators onto the floating gate via a tunneling effect
• This introduces ‘wear’ in the dielectric materials, e.g. by trapping / detrapping of charges in the materials and other defects

• With more and more program and erase activity, ‘wear’ decreases the isolating properties of the dielectric and changes its tunneling effectiveness, ultimately resulting in:
  • Incorrect reads from the NAND cell after relatively short retention times
  • Program / erase failures
SSD-Level Endurance: Controller and NAND Over-Provisioning

**NAND Components**

- [Image of NAND components]

**Controller**

- [Image of controller]

**SSD**

- [Image of SSD]

**Endurance Metric:**
- Number of P/E Cycles

**Typical Conditions:**
- Controller ECC Capability
- Data Retention

**JEDEC Test Conditions**

**Controller & drive design implement endurance management:**
- ECC
- Wear Leveling
- Data Refresh
- Data Redundancy
- Over-Provisioning
- Dynamic Read Trim / Read-Retry and/or Dynamic Write Trimming

**Endurance Metric:**
- Amount of Data Written to the SSD

**Endurance Conditions:**
- Write workload definition
- Effective Over-Provisioning / Drive ‘Fill-Level’
- Data Retention
- Drive UBER
- Date Entropy (Compressibility)

© 2012 HGST, a Western Digital company
SSD Write Endurance – A Definition

• Random Drive Writes per Day for 5 Years (DW/D)
  • ‘Drive Write’: amount of host write data equivalent of the drive’s capacity; in GB
• An equivalent metric is ‘Lifetime Random PB Written’ (PBW):
  • \[ \text{Endurance in PBW} = \text{Endurance in DW/D} \times \text{Drive Capacity in PB} \times 365 \text{ Days} \times 5 \text{ Years} \]

• Conditions are vendor-specific, example for HGST’s SSDs:
  • Random workload definition: IO size of 4KB or 8KB, 4K-aligned, full-volume random (access is uniformly distributed across the full drive volume)
  • Drive is 100% full – all LBA used, no LBA Trimmed / Unmapped
  • Data is completely random, e.g. not compressible
  • End-of-life, power-off data retention of 3 months in 40DegC storage temperature
  • UBER of 1 block in 10E16 Bytes read
  • Allowed performance degradation at end of rated product life <10%

*The definition of the endurance specification varies across vendors – you need to make sure you are comparing equivalent definitions*
Random vs. Sequential Write Endurance

- Due to random write amplification, sequential write endurance is typically higher than random write endurance.

- The larger the random WA of a particular SSD design, the larger the difference between the pure random and the pure sequential write endurance.

- But keep in mind that sequential write throughput is typically also higher than random write throughput, so endurance can be consumed more quickly as well.

Note: Unless noted otherwise, ‘Write Endurance’ refers to random write endurance in this presentation.
Real Life Write Workload Considerations

Less Write Endurance

- Not 100% random, some sequential
- Not 100% full volume, some parts of drive written more often than others
- Drive not 100% full – some LB never written and/or trimmed/unmapped

More Write Endurance

Workload used for HGST’s Write Endurance Specification

Depending on the actual write workloads, observed endurance can be significantly higher than HGST’s specification
‘Time to Wear-Out’

\[
\text{Time-to-Wear Out (Years)} = \frac{(\text{Endurance in DW/D} \times \text{Drive Capacity in GB} \times 5 \times 1000)}{\text{(Average Write Throughput in MB/s} \times 60 \times 60 \times 24)}
\]

<table>
<thead>
<tr>
<th>HGST Ultrastar SSD400S.B</th>
<th>Drive Capacity (GB)</th>
<th>Random Write Endurance (DW/D)</th>
<th>Max Sustained Random Write TP (MB/s)</th>
<th>Time-to-Wear-Out @ 100% Max Rand Write TP (Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>50</td>
<td>104</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>50</td>
<td>104</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>50</td>
<td>104</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HGST Ultrastar SSD400M</th>
<th>Drive Capacity (GB)</th>
<th>Random Write Endurance (DW/D)</th>
<th>Max Sustained Random Write TP (MB/s)</th>
<th>Time-to-Wear-Out @ 100% Max Rand Write TP (Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
<td>10</td>
<td>98</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>10</td>
<td>98</td>
<td>2.4</td>
</tr>
</tbody>
</table>

‘Worst-case’ numbers for ‘time-to-wear-out’ are typically a poor approximation of real-life Enterprise SSD use.
How Much Write Endurance is Enough?

‘Worst Case’
- 100% Write Mix & Max Write Throughput (i.e. 100% very high QD)
- 100% Duty Cycle
- HGST Workload Spec
  - 100% Random
  - Full-Volume Random
  - Drive 100% Full

‘Typical Case’
- Avg. of 30% of Max Write Throughput
- 80% Duty Cycle
- Typical workload:
  - Some sequential writes
  - Some write data locality
  - Drive not 100% full – TRIM/UNMAP

‘Time-to-Wear-Out’ Multiplier
- ‘Worst Case’
  - 3.3x
- ‘Typical Case’
  - 1.25x
  - 1.25x
  - 5.2x

‘Time-to-Wear Out’ Examples: HGST Ultrastar SSD400M
200GB: 1.2 Years
400GB: 2.4 Years
~ 6.3 Years
~ 12.5 Years
End of Rated Product Life – What to Expect

• Depends on SSD product design: One or more of the stated specifications will not be met any longer if SSD is used beyond the rated product life:
  • Power-off data retention
  • Read UBER
  • Write performance
    • As more NAND is retired due to program / erase failures, effective over-provisioning is reduced
  • Read performance
    • E.g. as more read ‘re-try’s are needed

• SSD is not likely to fail catastrophically immediately

Use SSD SMART statistics to monitor rated life use
Qualifying Drive Endurance – Best Practices

- Qualifying the SSD endurance specification is a time-consuming and costly effort
- Need to test large populations in order to get statically relevant data, especially when validating ‘rare event’ specifications like UBER
- Need to validate the actual end-of-life behavior of the SSD
  - Utilize specially configured, very small capacity SSDs to bring the SSD to wear-out point in ~2 months and validate all specifications
  - Also ensures end-of-life drive firmware paths are executed and tested thoroughly

Leading SSD vendors have built significant experience and know-how related to endurance validation testing over time
Agenda

SSD Market Context

Managing SSD Endurance

SSD Performance Considerations

SAS SSDs – Enterprise SSD Building Block
Throughput vs. Latency

Avg latency in ms = \(\frac{1}{\text{IOPS}} \times QD \times 1000\)

4KB Random Read QD32

QD1 to 8: ‘Ideal’ scaling of TP with QD
= very little increase in avg. latency

QD64 -> 128: No more TP gains = doubling in avg. latency

QD=1  2  4  8  16  32  64  128

IOPS

0  10,000  20,000  30,000  40,000  50,000  60,000

Avg. Latency (ms)

HGST SAS SSD
Performance Consistency – Command Completion Times

Narrow distributions of CCTs result in more consistent and deterministic performance.

4KB Random Read QD32 - Response Time

HGST SAS SSD

Avg.

# of commands

2012 Storage Developer Conference. © 2012 HGST, a Western Digital company
Agenda

SSD Market Context

Managing SSD Endurance

SSD Performance Considerations

SAS SSDs – Enterprise SSD Building Block
## Enterprise SSD – Interface Choices

For a given internal or external Enterprise storage system, numerous factors need to be considered to chose the most appropriate SSD interface.

<table>
<thead>
<tr>
<th>Interface</th>
<th>SATA</th>
<th>SAS</th>
<th>PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Command Set</td>
<td>ATA</td>
<td>SCSI</td>
<td>Proprietary or NVM Express or SCSI-over-PCIe</td>
</tr>
<tr>
<td>Main Form Factor</td>
<td>2.5”, 1.8”</td>
<td>2.5”</td>
<td>2.5”, Add-In Card</td>
</tr>
<tr>
<td>Max Device Power</td>
<td>9W Typical</td>
<td>9W Dual Port / 25W MultiLink SAS</td>
<td>25W</td>
</tr>
<tr>
<td>Transport Bandwidth</td>
<td>6 Gb / Port</td>
<td>6Gb / Port -&gt; 12Gb / Port</td>
<td>4Gb / Lane -&gt; 8Gb / Lane</td>
</tr>
<tr>
<td>Interface Configurations</td>
<td>Single Port</td>
<td>Dual Port / MultiLink SAS Four Ports</td>
<td>Four / Eight Lanes</td>
</tr>
<tr>
<td>Standardization</td>
<td>INCITS / SATA-IO</td>
<td>INCITS / STA</td>
<td>Vendor Specific; NVM Express Group, INCITS / STA; PCI-SIG</td>
</tr>
</tbody>
</table>
SAS SSDs enjoy the Enterprise maturity and support of the well-established SAS eco-system

### Maturity & Interoperability
‘Drop-In’ support in all major Enterprise system environments

### Scalability
Scales up to hundreds of drives with multi-port controllers & expanders

### High-Availability
Dual-port drives, T10 DIF, hot-plug support, cost-effective redundancy options using RAID controllers

### Technology Roadmap
12Gb SAS and Multi-Link SAS

### Standardization & Industry Support
Track record of effective standardization in INCITS; broad set of industry offerings
## PCIe as Drive Interface – Key Industry Initiatives

<table>
<thead>
<tr>
<th>Primary Target</th>
<th>Enterprise SSD</th>
<th>Enterprise SSD</th>
<th>Client SSD &amp; Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Host Command Interface</strong></td>
<td>SCSI (w/ SOP &amp; PQI)</td>
<td>NVMe (New)</td>
<td>ATA (via AHCI), followed by NVMe</td>
</tr>
<tr>
<td><strong>Transport</strong></td>
<td>PCIe x4</td>
<td>PCIe x4/x8</td>
<td>PCIe x2</td>
</tr>
<tr>
<td><strong>Form Factor</strong></td>
<td>2.5” Drive SFF Edge Card</td>
<td>2.5” Drive *, PCIe SIOM</td>
<td>2.5”, 1.8”, mSATA, etc.</td>
</tr>
<tr>
<td><strong>Connectors</strong></td>
<td>SFF 8639</td>
<td>SFF 8639, PCIe Edge Card</td>
<td>SATA-IO CabCon / SFF</td>
</tr>
<tr>
<td><strong>Standardization</strong></td>
<td>T10 &amp; STA</td>
<td>NVMe Group</td>
<td>SATA-IO</td>
</tr>
<tr>
<td><strong>OEM Endorsement</strong></td>
<td>HP, IBM, most STA members?</td>
<td>Dell, EMC, NetApp, Oracle, Cisco</td>
<td>TBD</td>
</tr>
<tr>
<td><strong>First SSD Products</strong></td>
<td>2013/2014</td>
<td>2013</td>
<td>2013/2014</td>
</tr>
</tbody>
</table>

2012 Storage Developer Conference.
SSD – Enterprise System Fit

Application needs associated with certain Enterprise system segments typically lead to an SSD product preference

<table>
<thead>
<tr>
<th>Storage System SAN &amp; NAS</th>
<th>Server App / DB</th>
<th>Blade Server</th>
<th>Cloud/ Web 2.0 Infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>🟢</td>
<td>🟢</td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>No Port Redundancy, No T10 DIF</td>
<td>Limited Interface Bandwidth at 6Gb, No T10 DIF</td>
<td>Limited Interface Bandwidth at 6Gb</td>
<td></td>
</tr>
<tr>
<td>SAS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>🟢 🟢 🟢</td>
<td>🟢</td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>PCIe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>🟢</td>
<td>🟢</td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>No Good Scaling Option</td>
<td>Limited RAID Options</td>
<td>Limited Interface Bandwidth at 6Gb</td>
<td>🟢 🟢 🟢 🟢 🟢 🟢</td>
</tr>
</tbody>
</table>

2012 Storage Developer Conference.