SSD Benchmarking at CERN

Liviu Vâlsan

HEPiX Fall 2014 Workshop, 16th of Oct 2014
SSD Benchmarking at CERN

- Technology overview
- SSD performance
  - Synthetic benchmarks
  - Simulations of Real Work applications
- SSD endurance
- Hurdles for SSD adoption
- Next steps
- Conclusion
SATA: A 10+ Years Old Interface for HDDs

- **2003**: SATA revision 1.0 - 1.5 Gbit/s - **150 MB/s**
- **2004**: SATA revision 2.0 - 3 Gbit/s - **300 MB/s**
- **2008**: SATA revision 3.0 - 6 Gbit/s - **600 MB/s**
- **2011**: SATA revision 3.1
  - No speed increase
  - Just extra form factors and features (TRIM)
- **2013**: SATA revision 3.2 (**SATA Express**) - 16 Gbit/s - **1969 MB/s**
NVMe Allows Standardization of PCIe SSDs

- A specification for accessing PCIe attached SSDs
- PCIe removes controller latency
- NVMe reduces software latency
- OS support
  - Linux support since kernel 3.3, stable since kernel 3.10
  - RHEL / SL 6.5, RHEL / CentOS / SL 7
  - Microsoft Windows Server 2012 R2, Windows 8.1
  - UEFI, QEMU, FreeBSD, Solaris
SATA Express

- Defines form factors / connectors that support either SATA or PCIe based drives
- Does not define the software interface
  - AHCI or NVMe can be used
- Exposes multiple PCIe lanes and two SATA gen 3.0 6 Gbit/s ports through the same host-side SATA Express connector
- Exposed PCIe lanes provide a pure PCIe connection to the SSD, without any additional layers of abstraction
Two Main Connectors

SFF-8639

Primary Side (closest to drive edge)
- Power, SATA Express sideband, P15-P1 (15 Pins)
- PCIe Lanes 3-1, Sideband E39-17 (23 Pins)
- SAS/SATA Express 2nd Port S14-S8 (7 Pins)
- PCIe Lane 0, RefClk E16-E7 (10 Pins)
- PCI Sideband E6-E1 (6 Pins)
- SAS/SATA/SATA Express 1st Port S7-S1 (7 Pins)

SATA Express
- Power
- SATA/PCIe Lane 0
- PCIe Lane 1
<table>
<thead>
<tr>
<th></th>
<th>AHCI</th>
<th>NVMe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum queue depth</td>
<td>1 command queue 32 commands per queue</td>
<td>65536 queues 65536 commands per queue</td>
</tr>
<tr>
<td>Uncacheable register accesses</td>
<td>4 per command - 8000 cycles ~ 2.5 μs</td>
<td>0 per command</td>
</tr>
<tr>
<td>(each consumes 2000 CPU cycles)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSI-X and interrupt steering</td>
<td>single interrupt no steering</td>
<td>2048 MSI-X interrupts</td>
</tr>
<tr>
<td>Parallelism and multiple threads</td>
<td>requires synchronization lock to issue a command</td>
<td>no locking</td>
</tr>
<tr>
<td>Efficiency for 4 KB commands</td>
<td>command parameters require two serialized host DRAM fetches</td>
<td>gets command parameters in one 64 Bytes fetch</td>
</tr>
</tbody>
</table>
# List of Benchmarked SSDs

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Family</th>
<th>Available capacities (GB)</th>
<th>Tested capacities (GB)</th>
<th>Interface</th>
<th>Flash Type</th>
<th>Endurance (DWPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>DC S3500</td>
<td>80, 120, 160, 240, 300, 480, 600, 800</td>
<td>240, 480</td>
<td>SATA rev. 3.0</td>
<td>20 nm MLC</td>
<td>0.3</td>
</tr>
<tr>
<td>Intel</td>
<td>DC S3700</td>
<td>100, 200, 400, 800</td>
<td>200, 800</td>
<td>SATA rev. 3.0</td>
<td>25 nm MLC</td>
<td>10</td>
</tr>
<tr>
<td>Intel</td>
<td>DC P3600</td>
<td>400, 800, 1200, 1600, 2000</td>
<td>400</td>
<td>PCIe gen 3 x4</td>
<td>20 nm MLC</td>
<td>3</td>
</tr>
<tr>
<td>Intel</td>
<td>DC P3700</td>
<td>400, 800, 1600, 2000</td>
<td>800</td>
<td>PCIe gen 3 x4</td>
<td>20 nm MLC</td>
<td>10</td>
</tr>
<tr>
<td>Intel</td>
<td>X25-E</td>
<td>32, 64</td>
<td>64</td>
<td>SATA rev. 2.0</td>
<td>50 nm SLC</td>
<td>18</td>
</tr>
<tr>
<td>Samsung</td>
<td>845DC Evo</td>
<td>240, 480, 960</td>
<td>240, 960</td>
<td>SATA rev. 3.0</td>
<td>19 nm TLC</td>
<td>0.35</td>
</tr>
<tr>
<td>Samsung</td>
<td>845DC Pro</td>
<td>400, 800</td>
<td>400</td>
<td>SATA rev. 3.0</td>
<td>40 nm MLC V-NAND</td>
<td>10</td>
</tr>
<tr>
<td>Samsung</td>
<td>SM843T</td>
<td>120, 240, 480, 960</td>
<td>240, 480</td>
<td>SATA rev. 3.0</td>
<td>20 nm MLC</td>
<td>2</td>
</tr>
<tr>
<td>Samsung</td>
<td>PM853T</td>
<td>240, 480, 960</td>
<td>240</td>
<td>SATA rev. 3.0</td>
<td>19 nm TLC</td>
<td>0.3</td>
</tr>
<tr>
<td>OCZ</td>
<td>Vertex 3</td>
<td>60, 90, 120, 240, 480</td>
<td>240</td>
<td>SATA rev. 3.0</td>
<td>25 nm MLC</td>
<td>0.3</td>
</tr>
</tbody>
</table>
4 KB Random Mixed 70% Read / 30% Write

Sustained 4 KB Random Mixed 70% Reads / 30% Writes Performance by # of Threads

# threads / outstanding IOs

IOPS
Sustained 4 KB Random Read Performance

Sustained 4 KB Random Read Performance by # of Threads

# threads / outstanding IOs

- Intel DC P3600 400 GB PCIe card HHHL
- Intel DC P3700 800 GB 2.5"
- Intel DC P3700 800 GB PCIe card HHHL
- Intel DC S3500 240 GB 2.5"
- Intel DC S3500 480 GB 2.5"
- Intel DC S3700 200 GB 2.5"
- Intel DC S3700 800 GB 2.5"
- Intel X25-E 64 GB 2.5"
- OCZ Vertex 3 240 GB 2.5"
- Samsung 845DC Evo 240 GB 2.5"
- Samsung 845DC Evo 960 GB 2.5"
- Samsung 845DC Pro 400 GB 2.5"
- Samsung PM853T 240 GB 2.5"
- Samsung SM843T 240 GB 2.5"
- Samsung SM843T 480 GB 2.5"
Sustained 4 KB Random Write Performance

Sustained 4 KB Random Write Performance by # of Threads

IOPS
Sustained 4 KB Random Write Latencies

Sustained 4 KB Random Write Latencies by # of Threads

Latency, µs

# threads / outstanding IOs

- Intel DC P3600 400 GB PCIe card HHHL
- Intel DC P3700 800 GB 2.5"
- Intel DC P3700 800 GB PCIe card HHHL
- Intel DC S3500 240 GB 2.5"
- Intel DC S3500 480 GB 2.5"
- Intel DC S3700 200 GB 2.5"
- Intel DC S3700 800 GB 2.5"
- Intel X25-E 64 GB 2.5"
- OCZ Vertex 3 240 GB 2.5"
- Samsung 845DC Evo 240 GB 2.5"
- Samsung 845DC Evo 960 GB 2.5"
- Samsung 845DC Pro 400 GB 2.5"
- Samsung PM853T 240 GB 2.5"
- Samsung SM843T 240 GB 2.5"
- Samsung SM843T 480 GB 2.5"
Performance Stability -- 4 KB Random Mixed 70% Read / 30% Write

Performance Stability -- Sustained 4 KB Random Mixed 70% Reads / 30% Reads Performance
Real-Work Application Use Cases

Large Blocks Analytics Engine

- Simulates a large-block analytics engine that is capable of streaming in data at a very high rate, sequentially scanning with multiple readers and a low (but non zero) update rate.
  - Block size: 128 KB
  - Number of jobs: 1
  - Writes: 10%
  - IO depth: 129
  - Access pattern: sequential writes
Big Block

- Simulates a large-block, aggregated checkpointing application where sequential checkpoint writes from many different systems in a HPC cluster are aggregated at a single node (turning those nice sequential streams into an effectively random write setup).
  - Block size: 512 KB
  - Number of jobs: 16
  - Writes: 100%
  - IO depth: 16
Throughput / USD * Capacity * Endurance Using Simulated Real World Applications

Throughput / USD * Capacity * Endurance Using Simulated Real World Applications (MB/s / USD * Capacity * Endurance)
Checkpointing

- Similar to Big Block, but this time with small, 4K checkpoint chunks.
- Block size: 4 KB
- Number of jobs: 16
- Writes: 100%
- IO depth: 16
- Access pattern: random writes
DB 8 KB

- Simulates a DB backend which uses 8 KB pages, without any logging or think time between reads/writes. This test stresses the IO subsystem without taking into account the locking, think times, etc. of a real DB.
- Block size: 8 KB
- Number of jobs: 8
- Writes: 40%
- IO depth: 4
OLTP

- Simulates an OLTP database pattern where the DB logs and data files are stored on the same SSD.
- Innosim is an InnoDB IO simulator that tests the disk IO capacity. Innosim mimics the workflow of a real DB (i.e. log writes are included and block transaction completions and data file updates).
- XFS file system used with a 4 KB block size
IOPS / USD * Capacity Using Simulated Real Work Applications
Performance / USD * Capacity Using Simulated Real World Applications (IOPS / USD * Capacity)
IOPS / USD * Endurance Using Simulated Real Work Applications

Performance / USD * Capacity * Endurance Using Simulated Real World Applications (IOPS / USD * Capacity * Endurance)
## SSD Endurance

### Intel S.M.A.R.T. Attributes

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Interpretation</th>
</tr>
</thead>
</table>
| 233 | Media_Wearout_Indicator     | **Normalized value:** reports the number of cycles the NAND media has undergone. Declines linearly from 100 to 1 as the average erase cycle count increases from 0 to the maximum rated cycles. Once the normalized value reaches 1, the number will not decrease, although it is likely that significant additional wear can be put on the device.  
**Raw value:** always 0. |
| 241 | Host_Writes_32MiB           | **Normalized value:** always 100.  
**Raw value:** reports the total number of sectors written by the host system. The raw value is increased by 1 for every 65,536 sectors (32MB) written by the host. |
| 226 | Workld_Media_Wear_Indic     | **Normalized value:** always 100.  
**Raw value:** measures the wear seen by the SSD (since reset of the workload timer, attribute E4h), as a percentage of the maximum rated cycles. Divide the raw value by 1024 to derive the percentage with 3 decimal points. |
| 228 | Workload_Minutes            | **Normalized value:** always 100.  
**Raw value:** measures the elapsed time (number of minutes since starting this workload timer). |
## SSD Endurance

### Samsung S.M.A.R.T. Attributes

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Interpretation</th>
</tr>
</thead>
</table>
| 177| Wear_Leveling_Count   | **Normalized value:** reports the number of cycles the NAND media has undergone. Declines linearly from 99 to 1 as the average erase cycle count increases from 0 to the maximum rated cycles. Once the normalized value reaches 1, the number will not decrease, although it is likely that significant additional wear can be put on the device.  
**Raw value:** the total count of P/E cycles. |
| 241| Total_LBAs_Written    | **Normalized value:** always 100.  
**Raw value:** the total size of all LBAs (Logical Block Address) required for all of the write requests sent to the SSD from the OS. To calculate the total size (in Bytes), multiply the raw value of this attribute by 512. |
Hurdles for SSD adoption

- Still a significant price difference compared to HDDs
- Still a gap in capacities compared to HDDs
- Almost no servers on the market with SATA Express / SFF-8639
- Currently no NVMe support in smartmontools, hdparm
- Many servers are still only SATA revision 2.0
- Ancient smartmontools version in SL 6 makes monitoring challenging
4 KB Random Mixed 70% Read / 30% Write - SATA 2 vs SATA 3

Sustained 4 KB Random Mixed 70% Reads / 30% Writes Performance by # of Threads

IOps

# threads / outstanding IOs

Samsung PM853T 240 GB 2.5" SATA 2.0
Samsung PM853T 240 GB 2.5" SATA 3.0
Samsung SM843T 240 GB 2.5" SATA 2.0
Samsung SM843T 240 GB 2.5" SATA 3.0
Multi-Threaded Random Read Bandwidth - SATA 2 vs SATA 3

Multi-Threaded Random Read Bandwidth SATA 2 vs SATA 3

Block Size (bytes)
Bogus smartmontools attribute names on SL 6

smartctl 5.43 2012-06-30 r3573 [x86_64-linux-3.14.5-1.el6.elrepo.x86_64] (local build)

Vendor Specific SMART Attributes with Thresholds:

<table>
<thead>
<tr>
<th>ID#</th>
<th>ATTRIBUTE_NAME</th>
<th>FLAG</th>
<th>VALUE</th>
<th>WORST</th>
<th>THRESH</th>
<th>TYPE</th>
<th>UPDATED</th>
<th>WHEN_FAILED</th>
<th>RAW_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>170</td>
<td>Unknown_Attribute</td>
<td>0x0033</td>
<td>100</td>
<td>100</td>
<td>010</td>
<td>Pre-fail</td>
<td>Always</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>171</td>
<td>Unknown_Attribute</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>172</td>
<td>Unknown_Attribute</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>174</td>
<td>Unknown_Attribute</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>175</td>
<td>Program_Fail_Count_Chip</td>
<td>0x0033</td>
<td>100</td>
<td>100</td>
<td>010</td>
<td>Pre-fail</td>
<td>Always</td>
<td>-</td>
<td>82173035128</td>
</tr>
<tr>
<td>180</td>
<td>Runtime_Bad_Block</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>192</td>
<td>Power-Off_Retract_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>21 (Min/Max 15/29)</td>
</tr>
<tr>
<td>194</td>
<td>Temperature_Celsius</td>
<td>0x0022</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>21</td>
</tr>
<tr>
<td>199</td>
<td>UDMA_CRC_Error_Count</td>
<td>0x003e</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>225</td>
<td>Load_Cycle_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>284832</td>
</tr>
<tr>
<td>226</td>
<td>Load-in_Time</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>65535</td>
</tr>
<tr>
<td>227</td>
<td>Torq-amp_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>4294967295</td>
</tr>
<tr>
<td>228</td>
<td>Power-off_Retract_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>65535</td>
</tr>
<tr>
<td>232</td>
<td>Available_Reservd_Space</td>
<td>0x0033</td>
<td>100</td>
<td>100</td>
<td>010</td>
<td>Pre-fail</td>
<td>Always</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>233</td>
<td>Media_Wearout_Indicator</td>
<td>0x0032</td>
<td>099</td>
<td>099</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>234</td>
<td>Unknown_Attribute</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>241</td>
<td>Total_LBAs_Written</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>284832</td>
</tr>
<tr>
<td>242</td>
<td>Total_LBAs_Read</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td>-</td>
<td>486496</td>
</tr>
</tbody>
</table>
Recent smartmontools version

```
smartctl 6.3 2014-07-26 r3976 [x86_64-linux-3.14.5-1.el6.elrepo.x86_64] (local build)
Vendor Specific SMART Attributes with Thresholds:

<table>
<thead>
<tr>
<th>ID#</th>
<th>ATTRIBUTE_NAME</th>
<th>FLAG</th>
<th>VALUE</th>
<th>WORST</th>
<th>THRESH</th>
<th>TYPE</th>
<th>UPDATED</th>
<th>WHEN_FAILED</th>
<th>RAW_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>170</td>
<td>Available_Reservd_Space</td>
<td>0x0033</td>
<td>100</td>
<td>100</td>
<td>010</td>
<td>Pre-fail</td>
<td>Always</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>171</td>
<td>Program_Fail_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>172</td>
<td>Erase_Fail_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>174</td>
<td>Unsafe_Shutdown_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>175</td>
<td>Power_Loss_Cap_Test</td>
<td>0x0033</td>
<td>100</td>
<td>100</td>
<td>010</td>
<td>Pre-fail</td>
<td>Always</td>
<td></td>
<td>632 (19 8681)</td>
</tr>
<tr>
<td>183</td>
<td>SATA_Downshift_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>190</td>
<td>Temperature_Case</td>
<td>0x0022</td>
<td>080</td>
<td>075</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>20 (Min/Max 15/29)</td>
</tr>
<tr>
<td>192</td>
<td>Unsafe_Shutdown_Count</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>194</td>
<td>Temperature_Internal</td>
<td>0x0022</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>199</td>
<td>CRC_Error_Count</td>
<td>0x003e</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>225</td>
<td>Host_Writes_32MiB</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>284834</td>
</tr>
<tr>
<td>226</td>
<td>Workld_Media_Wear_Indic</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>65535</td>
</tr>
<tr>
<td>227</td>
<td>Workld_Host_Reads_Perc</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>4294967295</td>
</tr>
<tr>
<td>228</td>
<td>Workload_Minutes</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>65535</td>
</tr>
<tr>
<td>232</td>
<td>Available_Reservd_Space</td>
<td>0x0033</td>
<td>100</td>
<td>100</td>
<td>010</td>
<td>Pre-fail</td>
<td>Always</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>233</td>
<td>Media_Wearout_Indicator</td>
<td>0x0032</td>
<td>099</td>
<td>099</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>234</td>
<td>Thermal_Throttle</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>0/0</td>
</tr>
<tr>
<td>241</td>
<td>Host_Writes_32MiB</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>284834</td>
</tr>
<tr>
<td>242</td>
<td>Host_Reads_32MiB</td>
<td>0x0032</td>
<td>100</td>
<td>100</td>
<td>000</td>
<td>Old_age</td>
<td>Always</td>
<td></td>
<td>486499</td>
</tr>
</tbody>
</table>
```
Next Steps

- Long term reliability measurements
  - Performance degradation over time
- Power measurements
- Finer grained monitoring of maximum latencies and standard deviations
- Improve monitoring of SSDs
- Evaluation of filesystem performance
- Evaluation of consumer drives
Conclusions

- Know your requirements
  - Performance
  - Endurance
- Know the limits of available SSDs
- Samsung SSDs provide better performance than Intel ones
- NVMe - significant performance improvements
- Think of developing SSD aware / friendly software
- Monitor the health of the drives
One more thing...
Results available online at:

http://cern.ch/go/7x8W