Opportunities and Challenges of Using Solid State Drives in Large Scale Datacenters

Badriddine M. Khessib
Dileep Bhandarkar
Microsoft Corporation
Overview

* SSD Usage Model and Application Classes
* Cost Model
* Application Specific Endurance Model
* Data Retention in the Data Center and end-of-life failure model
* Other SSD requirements
## SSD by the numbers

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Size (GB)</th>
<th>Price ($)</th>
<th>Perf</th>
<th>$/GB</th>
<th>$/Perf</th>
<th>Watts</th>
<th>W/GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>4</td>
<td>143</td>
<td>1000000</td>
<td>35.75</td>
<td>0.000143</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>SSD (SLC)</td>
<td>120</td>
<td>1244</td>
<td>10000</td>
<td>10.37</td>
<td>0.1244</td>
<td>2</td>
<td>0.017</td>
</tr>
<tr>
<td>SSD (MLC)</td>
<td>160</td>
<td>480</td>
<td>10000</td>
<td>3.00</td>
<td>0.048</td>
<td>2</td>
<td>0.013</td>
</tr>
<tr>
<td>SAS (15K)</td>
<td>300</td>
<td>216</td>
<td>200</td>
<td>0.72</td>
<td>1.08</td>
<td>14</td>
<td>0.047</td>
</tr>
<tr>
<td>SAS (10K)</td>
<td>300</td>
<td>186</td>
<td>150</td>
<td>0.62</td>
<td>1.24</td>
<td>8</td>
<td>0.027</td>
</tr>
<tr>
<td>SAS (7.2K)</td>
<td>2000</td>
<td>293</td>
<td>100</td>
<td>0.15</td>
<td>2.93</td>
<td>5</td>
<td>0.003</td>
</tr>
<tr>
<td>SATA (7.2K)</td>
<td>2000</td>
<td>293</td>
<td>100</td>
<td>0.15</td>
<td>2.93</td>
<td>5</td>
<td>0.003</td>
</tr>
</tbody>
</table>
Storage Technologies by the $*

SSD Usage Models

* HDD Caching
  * Intermediate persistent cache between HDDs and memory
  * Caches hot HDD pages in SSD
  * Cost efficient, but ..
  * Require hardware/software support

* HDD replacement
  * Easy to implement but could be too expensive
  * We have cost model for this approach
The Cost Model
(VLDB 2010 TPCTC workshop)

- HDD: IO is expensive
  - \( \text{Cost}_{\text{HDD}} = \text{IOPS}_H \times \$$/\text{IOPS}_{\text{HDD}} + \text{Power}_{\text{HDD}} \times \$$/\text{Watt} \)

- SSD: GB is expensive
  - \( \text{Cost}_{\text{SSD}} = \text{GB} \times \$$/\text{GB}_{\text{SSD}} + \text{Power}_{\text{SSD}} \times \$$/\text{Watt} \)

- For SSD to be viable:
  - \( \text{Cost}_{\text{HDD}} > \text{Cost}_{\text{SSD}} \)
  - \( \text{IOD} \times \$$/\text{IOPS}_{\text{HDD}} + \text{PD}_\Delta \times \$$/\text{Watt} > \$$/\text{GB}_{\text{SSD}} \)
The Cost Model (cont.)

IOD * $/IOPS_{HDD} + P_{D\Delta} * $/Watt > $/GB_{SSD}

- IOD: IOPS/GB, workload dependent
- $/IOPS_{HDD}$: $1.24
- $/Watt$: $10
- $/GB_{SSD}$: $10.37$ SLC & $3$ MLC

Solve for IOD:
- IOD > 8.28 (SLC)
- IOD > 2.34 (MLC)
## SSD Usage and Application Classes

<table>
<thead>
<tr>
<th></th>
<th>HDD caching</th>
<th>HDD replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Commodity Systems</strong></td>
<td>• Map/Reduce&lt;br&gt;• File system&lt;br&gt;• ECN</td>
<td>• Key/Value Store&lt;br&gt;• Web Search</td>
</tr>
<tr>
<td><strong>Reliable Systems</strong></td>
<td>• Enterprise OLTP&lt;br&gt;• Enterprise DSS</td>
<td>• ?</td>
</tr>
</tbody>
</table>
SSD is Consumable Storage

- Apps have to monitor State of SSD
  - SMART attributes
  - OS error events
  - App-level Page Checksums
- Costing Changes:
  - In enterprise a disk (HDD or SSD) is expected to last 3-4 and should be under warranty for that duration
  - For SSD Media is not covered with Warranty
  - Extra cost for the end user.
The Cost Model (Revisited)

- HDD: IO is expensive
  - \( \text{Cost}_{\text{HDD}} = \text{IOPS}_{\text{HDD}} \times \text{$/IOPSHDD} + \text{Power}_{\text{HDD}} \times \text{$/Watt} \)

- SSD: GB is expensive
  - \( \text{Cost}_{\text{SSD}} = \text{GB}_{\text{SSD}} \times \text{EF} \times \text{$/GBSSD} + \text{Power}_{\text{SSD}} \times \text{$/Watt} \)

- EF (Endurance Factor):
  - \( \text{App 3-year Writes (GB)/SSD endurance} \)
  - \( \text{EF} \geq 1 \)
SSD Endurance

- No standard way of specifying endurance
  - Some provide a single number based on certain workload
  - Some provide sequential and random numbers

- All are inadequate
  - Ignore IO block sizes
  - Assume long retention period (1 year)
Example Measured Endurance

Application Specific SSD endurance

- Application write TB (3 years)
- Max Experimental Evaluation write TB
- Max Vendor Spec write TB

Write TB

OLTP-1
OLTP-2
Mail
Text Search
Decision Support

Write TB values for different applications and use cases.
Application Centric Endurance Model

SSD endurance Model

App IO Profile

App-specific Endurance
SSD Endurance Model

* Find random and sequential SSD endurance for most block sizes: 4KB, 8KB, 16KB, 32KB, 64KB, 128KB, ..., 1MB
  * We collect SSD SMART attributes while running the above write workloads
  * The end result is figuring the write amplification model of the SSD.

* Disk used: 160GB MLC
  * 15TB (45TB overprovisioning) random endurance
  * 380 TB sequential endurance
Based on attribute 226: 785 TB
Based on attribute 233: 743 TB
OEM spec: ~385 TB.
Based on attribute 226: 277 TB
Based on attribute 233: 246 TB
No OEM spec.
4 KB Random Write

4 KB Random Writes

* Based on attribute 226: 122 TB
* Based on attribute 233: 112 TB
* OEM spec: 15TB
Based on attribute 226: 120 TB
Based on attribute 233: 93 TB
No OEM spec
Workload IO Characterization

* We use Windows ETW infrastructure to collect Disk IO traces
* Wrote tools to process those traces and extract:
  * Read/Write distribution by block size
  * Randomness
  * IO density
<table>
<thead>
<tr>
<th>Request Size</th>
<th>Total</th>
<th>% Total</th>
<th>Reads</th>
<th>% Read</th>
<th>Writes</th>
<th>% Writes</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>35860203</td>
<td>100%</td>
<td>35860203</td>
<td>100%</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>256</td>
<td>84559</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
<td>84559</td>
<td>97.3%</td>
</tr>
<tr>
<td>4</td>
<td>1749</td>
<td>0.0%</td>
<td>39</td>
<td>0.0%</td>
<td>1710</td>
<td>2.0%</td>
</tr>
<tr>
<td>32</td>
<td>205</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
<td>205</td>
<td>0.2%</td>
</tr>
<tr>
<td>28</td>
<td>133</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
<td>133</td>
<td>0.2%</td>
</tr>
<tr>
<td>8</td>
<td>103</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
<td>103</td>
<td>0.1%</td>
</tr>
<tr>
<td>24</td>
<td>82</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
<td>82</td>
<td>0.1%</td>
</tr>
<tr>
<td>12</td>
<td>70</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
<td>70</td>
<td>0.1%</td>
</tr>
<tr>
<td>16</td>
<td>65</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
<td>65</td>
<td>0.1%</td>
</tr>
</tbody>
</table>
Example Measured Endurance

**Application Specific SSD endurance**

- **Application write TB (3 years)**
- **Max Experimental Evaluation write TB**
- **Max Vendor Spec write TB**

<table>
<thead>
<tr>
<th>Application</th>
<th>Write TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLTP-1</td>
<td>200</td>
</tr>
<tr>
<td>OLTP-2</td>
<td>150</td>
</tr>
<tr>
<td>Mail</td>
<td>300</td>
</tr>
<tr>
<td>Text Search</td>
<td>100</td>
</tr>
<tr>
<td>Decision Support</td>
<td>500</td>
</tr>
</tbody>
</table>
Data Retention in the Data Center

- Very minimal data retention requirements
- Days not weeks
  - Data is replicated across servers and across data centers
  - If a server is down for few hours, rebuild server
- Servers always on
- We want to use SSD post 100% Media wear
Can we push SSD beyond per-spec 100% media wear?
* Answer is yes, based on our reduced data retention
* We already collect all SMART attributes/OS events
* Need the right SMART counters to predict “end” of life
  * Correctable ECC errors
  * Free blocks/retired blocks.

But …
* Certain SSDs will disable writes at 100% media wear
* Others do not throttle writes but provide no mechanism of detecting true end-of-life
Other Disk Requirements

* SMART counters:
  * Must:
    * % media wear
    * Host writes (GB)
  * Like:
    * Free blocks/retired blocks
    * FTL writes
    * ECC corrections
  * No endurance or end of life write throttling
    * Need for guaranteed SLA
  * Secure Erase
Endurance specification are ineffective and useless for Cloud apps:
  * Proposed a new app-specific endurance model

Data retention requirement in the cloud is not strong (few hours – days max)
  * We need to go beyond 100% media wear to fully utilize the disk
    * Need some visibility into the health of the disk (ECC corrections, for example)

No throttling at any stage: we need a predictable performance to maintain our SLA

Rich set of SMART counters will help us monitor and manage SSDs
  * Standardizing counters will simplify our software