Reza Bacchus
Hewlett Packard
Options Chief Technologist
NVDIMM: Super-Charging the Data Center
Agenda

- Memory Hierarchy
- Latency trends
- NVDIMM Types
- Accelerating Applications
- The Evolution of NVDIMMs
Memory Hierarchy

- **CPU**
- **Registers**
- **Cache**
- **DRAM**
- **NVM**
- **NAND Flash**
- **Magnetic**

**Shortest Latency, Highest Cost**

**Load/Store**
(Application Direct Access)

**Block**
(Application Indirect Access)

**Longest Latency, Lowest Cost**

Each level is an accelerator for the next lower level
Latency Trends

- HDD: 60 ms in 1987, 4 ms in 2018
- SSD: 0.05 ms in 2018
- NVM: 0.5 us in 2018

15x faster in 20 years
100k x faster in 5 years

Timing is everything – Invest Now!
### NVDIMM Types

#### Type 1 (NVDIMM-N)
- **Access Methods** -> Load/Store, Emulated Block
- **Capacity** = DRAM (10’s GB)
- **Latency** = DRAM (10’s of nanoseconds)
- **Energy source for backup**

#### Type 3 (NVDIMM-F)
- **Supported** -> Emulated Block
- **Capacity** = NAND (100’s GB – 1’s TB)
- **Latency** = NAND (10’s of microseconds)

#### Type 4
- **Supported** -> Load/Store, Emulated Block
- **Capacity** = NVM (100’s GB – 1’s TB)
- **Latency** = NVM (100’s of nanoseconds)
Customer Value by Workload

Low Cost

High Capacity

Persistency

**In Memory Database:** Journaling, reduced recovery time, Ex-large tables

**Traditional Database:** Log acceleration by write combining and caching

**Enterprise Storage:** Tiering and caching without an auxiliary power source

**Virtualization:** Higher VM consolidation with greater memory density

**High-Performance Computing:** Check point acceleration and/or elimination

**Other:** Object stores, unstructured data management, Genomics, archiving

NVDIMMs Accelerate a Rainbow of Applications
NVDIMM Acceleration

Baseline:
MySQL on FusionIO

Non-optimized:
Baseline + HP
NVDIMM + FusionIO
+ standard MySQL

Optimized:
Non-optimized + Flash-Aware MySQL

4x Acceleration in Throughput and Latency
The Evolution of NVDIMMs

- Origin as accelerator for cache on storage array controllers
  - Battery backed
- Type-1 accelerating database and HPC
  - Eliminating bottlenecks in journaling
- Type-3 accelerating High Frequency Trading
  - High IOPs
- Type-4 accelerating search, data manipulation, initialization
  - MicroController on NVDIMM
- And Beyond! FPGA on NVDIMM
  - Programmable on-the-fly
  - Accelerate dynamically changing workloads

NVDIMMs Accelerating Applications in the Future
Thank You!
RAM DISK Block IO Performance

Measured to the SNIA PTS

NVM Summit
January 20, 2015
St. Claire Hotel, San Jose CA

Eden Kim, CEO
Calypso Systems, Inc.
RAM Disk Block IO Performance

Non Volatile Memory – being advanced by NVMP TWG / NVDIMM SIG
  • Memory Mapped Load/Store
  • NVDIMM Block IO Type N / Type F
  • Faster than NAND Flash Based SSDs

In Memory Block IO Storage – data on a Linux RAM Disk
  • RAM Disk performance shows top end of Type N NVDIMM
  • RAM Disk Block IO performance is sensitive to settings
  • NVDIMM Block IO may approach RAM Disk performance

Linux RAM Disk is compared to Traditional Classes of NAND Flash SSD
Test Set Up

Hardware: PTS Reference Test Platform
- Intel S2600 COE Gen 3
- Dual Xeon 8 core, 3.2Ghz E5 2687W
- 32 GB (4GB x 8) DDR3 1600 ECC

Software:
- OS – CentOS 6.5
- Linux RAM Disk Block IO Driver 2.6.32-431.11.2.EL6.x86_64
- Test Software - Calypso CTS BE 1.9.216-eL6

Test Methodology: SNIA SSS Performance Test Specification v 1.1.1
### Storage Class

<table>
<thead>
<tr>
<th>Category</th>
<th>Device Type</th>
<th>Capacity</th>
<th>RND 4Kib 100% W</th>
<th>RND 4Kib 100% W</th>
<th>RND 4Kib 65:35 RW</th>
<th>RND 4Kib 100% R</th>
<th>SEQ 1024Kib 100% W</th>
<th>SEQ 1024Kib 100% R</th>
<th>RND 4Kib 100% W Ave</th>
<th>RND 4Kib 100% W Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HDD &amp; SSHD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>SSHD</td>
<td>7,200 RPM 2.5&quot; SATA Hybrid</td>
<td>500 GB</td>
<td>134</td>
<td>134</td>
<td>131</td>
<td>148</td>
<td>107 MB/s</td>
<td>103 MB/s</td>
<td>18.54 mSec</td>
</tr>
<tr>
<td>2</td>
<td>SAS HDD</td>
<td>15,000 RPM 3.5&quot; SAS HDD</td>
<td>80 GB</td>
<td>350</td>
<td>340</td>
<td>398</td>
<td>401</td>
<td>84 MB/s</td>
<td>90 MB/s</td>
<td>55.39 mSec</td>
</tr>
<tr>
<td><strong>CLIENT SSDs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>mSATA</td>
<td>mSATA 1.8&quot; MLC</td>
<td>128 GB</td>
<td>45,743</td>
<td>1,359</td>
<td>1,926</td>
<td>36,517</td>
<td>187 MB/s</td>
<td>533 MB/s</td>
<td>0.74 mSec</td>
</tr>
<tr>
<td>4</td>
<td>M.2 x2</td>
<td>M.2 x2 2280 MLC</td>
<td>512 GB</td>
<td>61,506</td>
<td>4,185</td>
<td>9,532</td>
<td>71,282</td>
<td>455 MB/s</td>
<td>535 MB/s</td>
<td>0.29 mSec</td>
</tr>
<tr>
<td>5</td>
<td>SATA Client</td>
<td>SATA 3 2.5&quot; MLC</td>
<td>200 GB</td>
<td>54,788</td>
<td>33,583</td>
<td>50,708</td>
<td>63,640</td>
<td>367 MB/s</td>
<td>480 MB/s</td>
<td>0.06 mSec</td>
</tr>
<tr>
<td><strong>ENTERPRISE SSDs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>SATA 6Gb/s</td>
<td>SATA 6Gb/s 2.5&quot; eMLC</td>
<td>800 GB</td>
<td>57,422</td>
<td>39,561</td>
<td>46,072</td>
<td>70,604</td>
<td>454 MB/s</td>
<td>504 MB/s</td>
<td>0.05 mSec</td>
</tr>
<tr>
<td>7</td>
<td>SAS 12Gb/s</td>
<td>SAS 12Gb/s 2.5&quot; MLC</td>
<td>800 GB</td>
<td>97,950</td>
<td>41,516</td>
<td>72,342</td>
<td>145,407</td>
<td>448 MB/s</td>
<td>973 MB/s</td>
<td>0.05 mSec</td>
</tr>
<tr>
<td>8</td>
<td>SFF 8639</td>
<td>SFF 8639 4 lane 2.5&quot; MLC</td>
<td>700 GB</td>
<td>149,512</td>
<td>44,872</td>
<td>166,002</td>
<td>397,564</td>
<td>564 MB/s</td>
<td>1,698 MB/s</td>
<td>0.01 mSec</td>
</tr>
<tr>
<td>9</td>
<td>PCIe 8 Lane</td>
<td>PCIe 8 Lane Edge Card MLC</td>
<td>1400 GB</td>
<td>159,926</td>
<td>87,419</td>
<td>236,227</td>
<td>742,674</td>
<td>614 MB/s</td>
<td>2,673 MB/s</td>
<td>0.01 mSec</td>
</tr>
</tbody>
</table>

All measurements taken on the RTP 3.0 CTS 6.5 Reference Test Platform pursuant to the SNIA PTS-E 1.1.

NOTE: Thread and Queue settings for PTS IOPS are T2Q16 for HDD/SSHD & Client SSDs and T4Q32 for Enterprise SSDs.
RAM Disk Application Demand Intensity

Effects on Performance Measurement
- NAND Flash: Pre Conditioning, Steady State, Write History
- RAM Disk: Outstanding IO, CPU cores, Memory Channels

Mapping RAM disks for Outstanding IOs (OIO) by Application Workload
- Vary Thread Count and Queue Depth
- Measure OIO by Response Time and IOPS (Demand Intensity)

Confidence Level Plot Compare (CLPC)
- Optimal OIO – Performance at different Thread Count x Queue Depth
- IOPS
- Response Time Quality of Service (QoS) – “5 9s” percentile response times
- Response Time Ceiling – Maximum response time allowed by application
Demand Intensity – RAM Disk: OIO x ART & IOPS

![Graph showing demand intensity for different RAM Disk settings with IOPS on the x-axis and average response time on the y-axis, indicating performance metrics for various workloads and configurations.]

- TC=32, QD=32,16,8,6,4,2,1
- TC=16, QD=32,16,8,6,4,2,1
- TC=6, QD=32,16,8,6,4,2,1
- TC=4, QD=32,16,8,6,4,2,1
- TC=2, QD=32,16,8,6,4,2,1
- TC=1, QD=32,16,8,6,4,2,1
- Mid: T8/Q1
- Max: T16/Q4

Min: T1Q1
Mid: T8Q1
Max: T16Q4
CLPC ‘QoS’ for a Single RAM Disk – RND 4KIB 100% Reads

Demand Intensity – Total Outstanding IOs
Comparison db OLTP: RAM Disk – PCIe x8 – SAS 12Gb/s

- RAM Disk
  - T32Q16 = OIO 512
  - Response Time (mSec) = 0.69

- PCIe 8
  - T16Q8 = OIO 128
  - Response Time (mSec) = 1.20

- SAS 12Gb/s
  - T16Q2 = OIO 32
  - Response Time (mSec) = 0.67
TAKE AWAYS

RAM Disk Block IO Performance is much higher than NAND Flash

NVDIMM RAM Block IO can approach the level of RAM Disk Block IO

RAM Disk Block IO Performance Depends on Settings

Applications can run much faster with RAM Disk and/or NVDIMM SSD

RAM Disk / NVDIMM SSD offer new Storage Tiering Opportunities
NVDIMMS in Enterprise Storage Arrays drive performance

Tom McKnight, Vice President of Hardware Platform @ Nimble Storage
Preface

- NVDIMMs combined with PCIe NTB’s have enabled Integrated Enterprise Storage Platforms to achieve significant performance improvements ( > 4X Write IOP latency improvement !! )

- Background:
  - An Integrated Enterprise / Fault Tolerant Storage Platform that implements software based RAID must preserve transient write data in the event of a power failure or hardware component failure. Data loss is NOT acceptable.
  - To insure fault tolerance the transient data must be replicated to the peer / standby controller before the write operation can be acknowledged. This inherent mirroring latency will heavily impact the systems maximum write performance ( typically measured in IOPS ).

**Lets compare the write performance of a legacy storage architecture against a new NVDIMM enabled architecture!**
Next Generation Integrated Enterprise Storage Array – Mirror NVData to / from NVDIMM via PCIe NTB
Action: External data write operation lands in Active DRAM buffer.
Legacy Integrated Enterprise Storage Array – Mirror NVData with Ethernet & PCIe NVRAM Cards

Action:
Copy data to local PCIe NVRAM card.
Legacy Integrated Enterprise Storage Array – Mirror NVData with Ethernet & PCIe NVRAM Cards

Action:
Send data via ethernet to standby controller.
Legacy Integrated Enterprise Storage Array – Mirror NVData with Ethernet & PCIe NVRAM Cards

Action: Standy controller copy’s data to it’s PCIe NVRAM card.
Legacy Integrated Enterprise Storage Array – Mirror NVData with Ethernet & PCIe NVRAM Cards

Action:
Standby sends “Ack” to Active controller via ethernet.
Legacy Integrated Enterprise Storage Array – Mirror NVData with Ethernet & PCIe NVRAM Cards

Action: Acknowledge the Host Write operation. Write is complete.
Legacy Integrated Enterprise Storage Array – Mirror NVData with Ethernet & PCIe NVRAM Cards

Action:
Write operation Acknowledged. IOP complete.
Next Generation Integrated Enterprise Storage Array – Mirror NVData to / from NVDIMM via PCIe NTB

**Action:**
External data write operation lands in Active DRAM buffer.
Next Generation Integrated Enterprise Storage Array – Mirror NVData to / from NVDIMM via PCIe NTB

Action: Memcopy data to NVDIMM from DRAM buffer.
Next Generation Integrated Enterprise Storage Array – Mirror NVData to / from NVDIMM via PCIe NTB

Action:
Active DMA transfer directly to Standby NVDIMM via PCIe NTB
Next Generation Integrated Enterprise Storage Array – Mirror NVData to / from NVDIMM via PCIe NTB

Action:
Write operation Acknowledged. IOP complete.
Conclusion

- NVDIMMs combined with PCIe NTB’s have enabled Integrated Enterprise Storage Platforms to achieve significant performance improvements ( > 4X Write IOP latency improvement !! )
And That’s All Folks.....

Thank You!