PCle SSD Panel Discussion

Moderator:
Eden Kim, Calypso Systems, Inc.

Panelists:
David Ridgeway, Fusion-io, Chuck Paridon, HP,
Jim Pappas, Intel, Marty Czekalski, Seagate, Swapna Yasarapu, STEC-Inc.
Agenda

Opening Remarks:
- Eden Kim, Calypso: Overview of Performance for different classes of storage

Panelist Remarks:
- David Ridgeway, Fusion-io: Software optimizations for NVM
- Chuck Paridon, HP: Case Study – Application workload characteristics that optimize PCIe SSD performance
- Jim Pappas, Intel: Next Generation Scalable NVM
- Marty Czekalski, Seagate: Challenges of Performance Metrics with new NVM Programming Models
- Swapna Yasrapu, STEC-Inc.: Mission – Reduce Application latency

Questions to Individual Panelists

Questions to Panel at large

Questions from Audience
## Moderator Remarks

### PCIe SSD Storage Performance Comparison

<table>
<thead>
<tr>
<th>Storage Class</th>
<th>Type</th>
<th>Device</th>
<th>Capacity &amp; WCE/WCD</th>
<th>FOB IOPS</th>
<th>IOPS (higher is better)</th>
<th>Throughput (larger is better)</th>
<th>Response Time (lower is better)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td></td>
<td></td>
<td>RND 4KB 100% W</td>
<td>RND 4KB 100% W</td>
<td>RND 4KB 65.35 RW</td>
<td>RND 4KB 100% R</td>
</tr>
<tr>
<td>HDD &amp; Hybrid</td>
<td>SSD 1</td>
<td>7,200 RPM 2.5” SATA Hybrid</td>
<td>500 GB WCD</td>
<td>125</td>
<td>147</td>
<td>150</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>SAS HDD 2</td>
<td>15,000 RPM 3.5” SAS HDD</td>
<td>80 GB WCD</td>
<td>350</td>
<td>340</td>
<td>398</td>
<td>401</td>
</tr>
<tr>
<td>Client SSDs</td>
<td>mSATA card 3</td>
<td>mSATA card 1.8” SSD</td>
<td>32 GB WCD</td>
<td>18,000</td>
<td>838</td>
<td>1,318</td>
<td>52,793</td>
</tr>
<tr>
<td></td>
<td>SATA SSD 4</td>
<td>SATA3 2.5” SSD</td>
<td>256 GB WCD</td>
<td>56,986</td>
<td>3,147</td>
<td>3,779</td>
<td>29,876</td>
</tr>
<tr>
<td></td>
<td>SATA SSD 5</td>
<td>SATA3 2.5” SSD</td>
<td>256 GB WCE</td>
<td>60,090</td>
<td>60,302</td>
<td>41,045</td>
<td>40,686</td>
</tr>
<tr>
<td>Enterprise SSDs</td>
<td>SAS SSD 6</td>
<td>Enterprise 2.5” SAS SSD</td>
<td>400 GB WCD</td>
<td>61,929</td>
<td>24,848</td>
<td>29,863</td>
<td>53,942</td>
</tr>
<tr>
<td></td>
<td>PCIe SSD 7</td>
<td>Server 2.5” PCIe SSD</td>
<td>320 GB WCD</td>
<td>133,580</td>
<td>73,008</td>
<td>53,797</td>
<td>54,327</td>
</tr>
<tr>
<td></td>
<td>PCIe Card 8</td>
<td>Server Edge Card PCIe SSD</td>
<td>700 GB WCD</td>
<td>417,469</td>
<td>202,929</td>
<td>411,390</td>
<td>684,284</td>
</tr>
</tbody>
</table>

*Figure A-1*  
All measurements taken by Calypso Systems, Inc. on the SSSI RTP/CTS test platform pursuant to PTS 1.1
Updates to Performance Comparison Chart in 2014

- M.2 form factor SATA Express 2/4 lane PCIe
- 12 Gb/s SAS
- SFF 8639 2.5” form factor 4 lane PCIe SSD
- PCIe Edge card on Gen 3 bus
- NVM DIMM SSD
- NVM Memory Channel Technology

Go to www.calypsotesters.com/news
Software Optimizations for Non-Volatile Memory
At Fusion-io, David Ridgeway is responsible for managing Fusion-io cloud and webscale ioMemory Flash products to deliver state of the art application acceleration design and products.

Prior to joining Fusion-io, Mr. Ridgeway served as Director Product Management for the 10G Ethernet and I/O Virtualization products at Emulex. David blends his knowledge in compute, network and storage products with his knowledge of the high-tech industry in bringing datacenter products to market. Mr. Ridgeway holds a BS Electrical Engineering Technology from Cogswell Polytechnic College and has over 25 years of experience with Emulex, PLX, Xilinx and GTE.
3 Contributions to the Community

Current OpenNVM Repositories

Flash-aware Linux swap
When working set size exceeds the capacity of DRAM, demand page from a flash-aware virtual memory subsystem.
[Repository] [Learn More]

Key-value interface to flash
Create NoSQL databases faster. Automate garbage collection of expired data.
[Repository] [Learn More]

Flash programming primitives
Use built-in characteristics of the Flash Translation Layer to perform journal-less updates (more performance and less flash wear = lower TCO)
[Repository] [Learn More]

https://opennvm.github.io
Atomic Writes – MySQL Example

Traditional MySQL Writes

MySQL with Atomic Writes
Flash Primitives: Sample Uses and Benefits

**Databases**

- **Transactional Atomicity:**
  Replace various workarounds implemented in database code to provide write atomicity
  - example: MySQL double-buffered writes

**Filesystems**

- **File Update Atomicity:**
  Replace various workarounds implemented in filesystem code to provide file/directory update atomicity
  - example: journaling

- **98% performance of raw writes**
  Smarter media now natively understands atomic updates, with no additional metadata overhead.

- **2x longer flash media life**
  Atomic Writes can increase the life of flash media up to 2x due to reduction in write-ahead-logging and double-write buffering.

- **50% less code in key modules**
  Atomic operations dramatically reduce application logic, such as journaling, built as work-arounds.
2-4x Latency Improvement on Percona Server

Sysbench 99% latency OLTP workload

Latency

Seconds 3600

XFS DoubleWrite

Atomic Writes
70% Transactions/sec Improvement on MariaDB Server

XtraDB 5.5.30 – Atomics | TPC-C - 2500 warehouses | 230GB data - 50GB buffer pool

New Order Transactions (10 sec)
OpenNVM, Standards, and SNIA NVM-Programming TWG

- Opennvm.github.io
  - Primitives API specifications, sample code
  - Linux swap kernel patch and benchmarking tools
  - Key-value interface API library, sample code, benchmark tools

- INCITS SCSI (T10) active standards proposals:
  - SBC-4 SPC-5 Atomic-Write
    http://www.t10.org/cgi-bin/ac.pl?t=d&f=11-229r6.pdf
  - SBC-4 SPC-5 Scattered writes, optionally atomic
    http://www.t10.org/cgi-bin/ac.pl?t=d&f=12-086r3.pdf
  - SBC-4 SPC-5 Gathered reads, optionally atomic
    http://www.t10.org/cgi-bin/ac.pl?t=d&f=12-087r3.pdf

- SNIA NVM-Programming TWG draft guide:
  http://snia.org/forums/ssssi/nvmp

- Apps Using OpenNVM technology:
Jim Pappas is the Director of Technology Initiatives in Intel’s Data Center Group. In this role, Jim is responsible to establish broad industry ecosystems that comply with new technologies in the areas of Enterprise I/O, Energy Efficient Computing, and Solid State Storage. Jim has founded, or served on several organizations in these areas including: PCI Special Interest Group, SNIA, InfiniBand Trade Association, Open Fabrics Alliance, The Green Grid, and several emerging initiatives in his newest focus area of Solid State Storage.

Mr. Pappas has previously been the Director of Technology Initiatives in Intel’s Desktop Products Group, and successfully led technologies such as AGP Graphics, DVD, IEEE 1394, Instantly Available PC, PCI, USB, and other advanced technologies for the Desktop PC.

Mr. Pappas has 31 years of experience in the computer industry. He has been granted eight U.S. patents in the areas of computer graphics and microprocessor technologies. He has spoken at major industry events including the Intel Developer’s Forum (IDF), WinHEC, Storage Networking World, PC Strategy, Microprocessor Forum, Consumer Electronics Show, Server I/O and the Applied Computing Conference. He holds a B.S.E.E. from the University of Massachusetts, Amherst, Massachusetts.
NVM Summit - PCIe Panel Discussion

Jim Pappas - jim@intel.com
Next Generation Scalable NVM

Resistive RAM NVM Options

<table>
<thead>
<tr>
<th>Family</th>
<th>Defining Switching Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase Change Memory</td>
<td>Energy (heat) converts material between crystalline (conductive) and amorphous (resistive) phases</td>
</tr>
<tr>
<td>Magnetic Tunnel Junction (MTJ)</td>
<td>Switching of magnetic resistive layer by spin-polarized electrons</td>
</tr>
<tr>
<td>Electrochemical Cells (ECM)</td>
<td>Formation / dissolution of “nano-bridge” by electrochemistry</td>
</tr>
<tr>
<td>Binary Oxide Filament Cells</td>
<td>Reversible filament formation by Oxidation-Reduction</td>
</tr>
</tbody>
</table>

Cross Point Array in Backend Layers \(\sim 4\lambda^2\) Cell

Scalable Resistive Memory Element

Wordlines

Memory Element

Selector Device

Bitlines

\(~ 1000x\) speed-up over NAND.
Opportunities with Next Generation NVM

Application to SSD IO Read Latency (us, QD=1, 4KB)

Next Generation NVM – Media is no longer the bottleneck
Opportunities with Next Generation NVM

Application to SSD IO Read Latency (us, QD=1, 4KB)

NAND MLC SATA 3 ONFI 2
NAND MLC SATA 3 ONFI 3
NAND MLC PCIe Gen 3 ONFI 3
Future NVM PCIe x4 Gen 3
Future NVM PCIe x4 Gen 3

Interconnect & Software Stack Dominate the Access Time
Next Generation Scalable NVM

Resistive RAM NVM Options

<table>
<thead>
<tr>
<th>Family</th>
<th>Defining Switching Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase Change Memory</td>
<td>Energy (heat) converts material between crystalline (conductive) and amorphous (resistive) phases</td>
</tr>
<tr>
<td>Magnetic Tunnel Junction (MTJ)</td>
<td>Switching of magnetic resistive layer by spin-polarized electrons</td>
</tr>
<tr>
<td>Electrochemical Cells (ECM)</td>
<td>Formation / dissolution of “nano-bridge” by electrochemistry</td>
</tr>
<tr>
<td>Binary Oxide Filament Cells</td>
<td>Reversible filament formation by Oxidation-Reduction</td>
</tr>
</tbody>
</table>

Cross Point Array in Backend Layers ~4λ² Cell

~ 1000x speed-up over NAND.
Opportunities with Next Generation NVM

Application to SSD IO Read Latency (us, QD=1, 4KB)

- NAND MLC SATA 3 ONFI 2
- NAND MLC SATA 3 ONFI 3
- NAND MLC PCIe Gen 3 ONFI 3

Existing SSD Latency Profile

Legend:
- NVM Tread
- NVM Xfer
- Misc SSD
- Link Xfer
- Platform + Adapter
- Software
Opportunities with Next Generation NVM

Application to SSD IO Read Latency (us, QD=1, 4KB)

Next Generation NVM – Media is no longer the bottleneck
Opportunities with Next Generation NVM

Interconnect & Software Stack Dominate the Access Time
Opportunities with Next Generation NVM

NVM Express/SCSI Express: Optimized storage interconnect & driver
SNIA NVM Programming TWG: Optimized system & application software
Chuck Paridon
Hewlett-Packard
Storage Performance Architect

Case Study: Application Workload Characteristics That Optimize PCIe SSD Performance
Chuck is a 27 year veteran of computer and storage subsystem performance analysis. He is responsible for the development and delivery of HP field performance training. A member of the SNIA SSI and GSI TWGS, and the Storage Performance Council, He holds a BS in ME and an MS in Computer Science.
Outline

- Optimal Application Workload Characteristics
  - Hot spots and Their Behavior
  - Measuring the Capacity and Intensity of Hot Spots
- Application of PCIe Storage to Optimize OLTP
  - Defining what portions of an OLTP workload are best optimized via PCIe SSD Storage
  - Cache Design Parameters
- Conclusion and Results of PCIe Cache Implementation
  - Increase IO performance via cache hits
Optimal Application Characteristics

- Asymmetric Regions of IO Demand “Hot Spots” within Subregions of the Application Address Space

- The Location of these Regions must Persist Sufficiently to Allow the Caching algorithms to Relocate them within the PCIe Cache

- The Relative Demand Intensity of these Regions must be Sufficiently high to take full Advantage of PCIe SSD Performance
Optimal Application Characteristics

Typical IO Density Curve for OLTP Applications Here we see that 90% of the IOS reside in about 7% of the Total Space. This workload is very well suited to take advantage of the high performance of PCIe storage.
From the Preceding Chart, we can Glean the Following:

- Based on relative IOPS, we see that 90% of the IO demand can be contained within 7% of the total space.

- By knowing the proposed tier performance density, and the total space used, we can calculate the application speedup of the PCIe cache enabled solution.
  
  - Consider the total space required to be 100 TB and the total proposed IO demand to be 58,600 IOPS using 15k RPM, 300 GB drives. (this was measured on 533 drives @ 110 IOPS /HDD)
  
  - Calculating the performance density of the PCIe tier (41,890 IOPS/1,400 GB) is 29.92 IOPS per GB *
  
  - Workload is assumed to be 8 KB 40/60 mix

* Device measurements courtesy of Eden Kim / Calypso Systems, Inc.
### Storage Performance Template: (Raid5)

- **Workload Array Limits**
- **200 GB SSD**
- **400 GB SSD**
- **800 GB SSD**
- **15k 300 GB**
- **15k 450 GB**
- **15k 600 GB**
- **10k 300 GB**
- **10k 450 GB**
- **10k 600 GB**
- **7.2k 1 TB**
- **7.2k 2 TB**

<table>
<thead>
<tr>
<th>Workload</th>
<th>Array Limits</th>
<th>200 GB SSD</th>
<th>400 GB SSD</th>
<th>800 GB SSD</th>
<th>15k 300 GB</th>
<th>15k 450 GB</th>
<th>15k 600 GB</th>
<th>10k 300 GB</th>
<th>10k 450 GB</th>
<th>10k 600 GB</th>
<th>7.2k 1 TB</th>
<th>7.2k 2 TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Reads (IOPS)</td>
<td>365,500</td>
<td>28.82 IOPs/GB</td>
<td>14.41 IOPs/GB</td>
<td>7.20 IOPs/GB</td>
<td>1.45 IOPs/GB</td>
<td>0.967 IOPs/GB</td>
<td>0.725 IOPs/GB</td>
<td>1.19 IOPs/GB</td>
<td>0.793 IOPs/GB</td>
<td>1.45 IOPs/GB</td>
<td>230 IOPs/TB</td>
<td>115 IOPs/TB</td>
</tr>
<tr>
<td>Random Writes (IOPS)</td>
<td>139,500</td>
<td>11.67 IOPs/GB</td>
<td>5.84 IOPs/GB</td>
<td>2.99 IOPs/GB</td>
<td>0.378 IOPs/GB</td>
<td>0.252 IOPs/GB</td>
<td>0.189 IOPs/GB</td>
<td>0.310 IOPs/GB</td>
<td>0.207 IOPs/GB</td>
<td>1.45 IOPs/GB</td>
<td>69.86 IOPs/TB</td>
<td>34.93 IOPs/TB</td>
</tr>
<tr>
<td>Rnd 70/30 Mix (IOPS)</td>
<td>256,040</td>
<td>15.32 IOPs/GB</td>
<td>7.66 IOPs/GB</td>
<td>3.83 IOPs/GB</td>
<td>0.794 IOPs/GB</td>
<td>0.529 IOPs/GB</td>
<td>0.397 IOPs/GB</td>
<td>0.651 IOPs/GB</td>
<td>0.433 IOPs/GB</td>
<td>1.45 IOPs/GB</td>
<td>125.2 IOPs/TB</td>
<td>62.60 IOPs/TB</td>
</tr>
<tr>
<td>Seq. Reads (MB/s)</td>
<td>14,660</td>
<td>0.304 MB/s/GB</td>
<td>0.304 MB/s/GB</td>
<td>0.304 MB/s/GB</td>
<td>0.203 MB/s/GB</td>
<td>0.152 MB/s/GB</td>
<td>0.249 MB/s/GB</td>
<td>0.166 MB/s/GB</td>
<td>0.304 MB/s/GB</td>
<td>84.82 MB/s/TB</td>
<td>42.41 MB/s/TB</td>
<td></td>
</tr>
<tr>
<td>Seq. Writes (MB/s)</td>
<td>10310</td>
<td>0.456 MB/s/GB</td>
<td>0.228 MB/s/GB</td>
<td>0.114 MB/s/GB</td>
<td>0.052 MB/s/GB</td>
<td>0.035 MB/s/GB</td>
<td>0.026 MB/s/GB</td>
<td>0.043 MB/s/GB</td>
<td>0.029 MB/s/GB</td>
<td>0.304 MB/s/GB</td>
<td>10.58 MB/s/TB</td>
<td>5.29 MB/s/TB</td>
</tr>
</tbody>
</table>
Implementing the Cache

- Let’s use the same 300 GB 15k RPM drives for the slower tier (only change one parameter at a time)
  - We also know we need 93% of 100TB, or roughly 93 TB of lower speed capacity.
  - This translates to $93,000 / 0.625 = 149$ TB raw = ~ 496 HDDs

- Interpolating for the 40/60 workload using the previous table, we get the delivered performance density of $93,000 * 0.586$ IOPS /GB = 54,500 IOPS
  - This is the contribution of the slower tier to the overall solution performance.

- So now, the performance of the new PCIe tiered solution is:
  - 209,200 IOPS Fast Tier
  - 54,500 IOPS Slower Tier
  - 263,700 IOPS Total
Proposed Solution IO Rate Comparison

- New Solution yields 263,700 IOPs
- Original Solution yielded 58,600 IOPS

Performance Increase = 4.5 x.
Thank you
Marty Czekalski
Seagate Technology
Interface and Emerging Architecture Program Manager
President SCSI Trade Association

"Challenges of Performance Metrics with New NVM Programming Models"
Marty Czekalski brings over thirty years of senior engineering management experience in advanced architecture development for Storage and IO subsystem design, ASIC, and Solid State Storage Systems.

He is currently Sr. Staff Program Manager within Seagate’s Emerging Business Organization.

Previous industry experience includes engineering management roles at Maxtor, Quantum and Digital Equipment Corporation.

Mr. Czekalski has participates in multiple interface standards committees and industry storage groups. He was a founding member of the Serial Attached SCSI Working Group during that lead to the development of Serial Attached SCSI. He currently serves as President and member of the Board of Directors of the SCSI Trade Association. Mr. Czekalski is also an active member of T10, SNIA, PCI-SIG and JEDEC.

Mr. Czekalski earned his MS degree in Electrical Engineering from the University of Maryland, and his BE degree in Electrical Engineering from the Stevens Institute of Technology.
Today’s Performance Metrics

- CPU
- Caches
- DRAM
- IO Subsystem
- Storage

Persistence Boundary

- Bandwidth
- Latency
- IOPs

- Persistent
- Durable
- Snap Shots
- Multi Access
- Access Control
- Error handling
- Management
- Security
Relative Cost vs Time Chart

- New NVM
- DRAM
- Flash
- HDD

Phase Change, Memristor, RRAM, etc.
PM Programming Models

Source: SNIA NVM Programming Model V1.0.0 Revision 5
PM Requires Rethinking Performance Metrics

- Bandwidth?
- Latency?
- IOPs?
- Access time?
- Cache flush?
- Wear leveling?
- Replication?

- Persistent
- Durable
- Snap Shots
- Multi Access
- Access Control
- Error handling
- Management
- Security
.httpClientRequest(2, 0, 0)

function request (url, method, data, config) { return httpRequest (url, method, data, config) }
As Director, SSD Product marketing HGST, Swapna Yasarapu is charged with defining and delivering next generation SSDs to enterprises. Swapna has been involved in defining and delivering state of the art SSDs since the early stages of SSD technology introduction to the market.

With a Masters in Electrical Engineering from University of California, Irvine specializing in Computer systems and software and MBA from Anderson School of Management, University of California, LA with emphasis in Marketing and Strategy, Swapna blends her technical knowledge with her knowledge of the hi-tech industry in bringing enterprise class products to market.

Swapna has over 15 years of in depth experience in hi-tech storage networks with responsibilities spanning from ASIC development, product development to managing hardware and software storage products through concept and development to production.
Mission: Reduce Application Latency
Fulfilment: Optimize the complete IO stack

Applications
Eg. Oracle Flash Cache

File System Optimization
Operating System optimization
Driver Stack

Modern Interface Protocols
(NVM Express, SCSI Express)

Non-volatile Media
(Flash/SSDs, Future NVM devices – PCRAM, MRAM etc)
Questions & Answers
Questions to Individual Panelists

Chuck Paridon, HP

Q1. What is the main risk of implementing PCIe SSD storage with mission critical applications?
Questions to Individual Panelists

Chuck Paridon, HP

Q2. What is the limiting factor in expanding PCIe SSD capacity in a server?
Questions to Individual Panelists

Marty Czekalski, Seagate

Q1. With PCIe SSDs positioned as a premium product, when if ever will adoption catch up with SATA SSD, which cost less and provide a nearly equivalent user experience in the majority of real world client use cases?
Marty Czekalski, Seagate

Q2. What are the issues holding back "true hot plug" (surprise insertion/removal) of PCIe storage devices?
Q1. The year is 2020 and you are Chief Storage Technologist at Unbelievable Storage Enterprise Resources – Deliverable User Devices and Extensions (USER-DUDE).

You have an unlimited budget and a mission to install a system-wide storage infrastructure for immediate deployment. Focusing on PCIe SSD extensions from the year 2014, what and how would you accomplish this mission? You may define the applications and markets for which your solution will address.

i.e. “What is the future for cool and super high performance storage equipment?”
Q2. What is the future of high performance PCIe SSD storage? What are the killer apps that will be developed to take advantage of emerging technologies such as NVM, PM and Storage Class Memories?

i.e. “What are the future apps that take advantage of cool and super high performance storage?”
Thank You
Questions to Individual Panelists

Jim Pappas, Intel

Q1. xxx

Q2. xxx
Questions to Individual Panelists

David Ridgeway, Fusion-io

Q1. xxx

Q2. xxx
Questions to Individual Panelists

Swapna Yasurupa

Q1. xxx

Q2. xxx