NVMe: Next Generation SSD Interface

Anil Vasudeva, President & Chief Analyst
IMEX Research
The material contained in this tutorial is copyrighted by the SNIA and author unless otherwise noted.

Member companies and individual members may use this material in presentations and literature under the following conditions:

- Any slide or slides used must be reproduced in their entirety without modification
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA Education Committee.

Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
Abstract

Computer architects dream of storage devices which can provide very high IOPs at minimal cost (IOPS/$/GB) using infinite cheap storage and instant access (low latency) for their applications/workloads.

Enterprise-Ready SSDs have started to fulfill that promise and being available in many formats from Servers based (DIMM Based - NVDIMM, SATA DIMM and PCIe based NVMe, SCSI Express – SAS/PCIe) and Storage (SAS/SATA and PCIe based Products). But it is only the recent advent of NVMe Interface for DRAM that holds the promise of surpassing NAND in performance and become the next standard being endorsed and promulgated by over 80+ vendors including top 15 major Server & Storage Suppliers.

The presentation delineates the recently emerged technologies, storage characteristics (performance, cost, reliability and endurance etc.), and the applications that benefit the most from the use of SSDs in enterprise storage systems, workloads optimization using new generation of controllers and automated smart-tiering by specific vertical-industries as well as the economics of SSDs usage using real market data.

Learning Objectives

The session provides a clear illustrative views of the state of interfaces developing in SSDs. Advances applicable for SSDs as caching devices for implementation in Server and Storage environments. This would allow the audience to be abreast of market and IT drivers (such as Virtualization, Cloud, Big Data etc) driving the SSD technologies for improvements in performance, cost, reliability and endurance etc..

overview of industry advances, specific applications that benefit the most from SSDs use, system implementation in enterprise storage systems allowing them to plan, implement and achieve stated benefits expected from using SSDs as a tiered storage specifically in OLTP/Database, Business Intelligence applications and cluster-based HPC workloads. The presentation illustrates how servers using NVMe interface in the Flash Storage stack can improve performance by 50% over PCIe based SCSI/SAS SSDs.
IT Industry’s Journey - Roadmap

Analytics – BI
Predictive Analytics - Unstructured Data
From Dashboards Visualization to Prediction Engines using Big Data.

Cloudization
On-Premises > Private Clouds > Public Clouds
DC to Cloud-Aware Infrast. & Apps. Cascade migration to SPs/Public Clouds.

Automation
Automatically Maintains Application SLAs
(Self-Configuration, Self-Healing©IMEX, Self-Acctg. Charges etc)

Virtualization
Pools Resources. Provisions, Optimizes, Monitors
Shuffles Resources to optimize Delivery of various Business Services

Integration/Consolidation
Integrate Physical Infrast./Blades to meet CAPSIMS ©IMEX
Cost, Availability, Performance, Scalability, Inter-operability, Manageability & Security

Standardization
Std. IT Infrastructure- Volume Economics HW/Syst SW
(Servers, Storage, Netwk Devices, System Software (OS, MW & Data Mgmt SW)

Source: Flash Industry Report IMEX Research ©2009-12
NVMe: The NextGen Interface for Solid State Storage
© 2012 Storage Networking Industry Association. All Rights Reserved.
Request for data from a remote client to an enterprise data center crosses a myriad of systems and devices. Key is identifying bottlenecks & improving performance.
Virtualized Cloud Infrastructure

Application’s SLA dictates the Resources Required to meet specific requirements of Availability, Performance, Cost, Security, Manageability etc.
Big Data Analytics – Next Frontier in IT

![Graph showing the relationship between transaction data and interaction data in the context of databases and analytics.](image)

Source: Flash Industry Report IMEX Research ©2009-12

© 2012 Storage Networking Industry Association. All Rights Reserved.
Market Segments by Apps/Workloads

*IOPS* for a required response time (ms)
*=(#Channels*Latency-1)

**OLTP**
**Transaction Processing**
**eCommerce**
**Data Warehousing**
**Business Intelligence**

**OLAP**

**(RAID - 0, 3)**

**(RAID - 1, 5, 6)**

**Scientific Computing**
**HPC**
**Imaging**
**Audio**
**Video**
**Web 2.0**

TP
HPC

Source: Flash Industry Report IMEX Research ©2009-12

NVMe: The NextGen Interface for Solid State Storage
© 2012 Storage Networking Industry Association. All Rights Reserved.
Anatomy of Data Access

Time taken by CPU, Memory, Network, Disk for a typical I/O Operation during a Data Access

<table>
<thead>
<tr>
<th>Action</th>
<th>Time (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 cache reference</td>
<td>0.5</td>
</tr>
<tr>
<td>Branch mispredict</td>
<td>5</td>
</tr>
<tr>
<td>L2 cache reference</td>
<td>7</td>
</tr>
<tr>
<td>Mutex lock/unlock</td>
<td>25</td>
</tr>
<tr>
<td>Main memory reference</td>
<td>100</td>
</tr>
<tr>
<td>Compress 1K bytes with Zippy</td>
<td>3,000</td>
</tr>
<tr>
<td>Send 2K bytes over 1 Gbps network</td>
<td>20,000</td>
</tr>
<tr>
<td>Read 1 MB sequentially from memory</td>
<td>250,000</td>
</tr>
<tr>
<td>Round trip within same datacenter</td>
<td>500,000</td>
</tr>
<tr>
<td>Disk seek</td>
<td>10,000,000</td>
</tr>
<tr>
<td>Read 1 MB sequentially from disk</td>
<td>20,000,000</td>
</tr>
<tr>
<td>Send packet CA-&gt;Netherlands-&gt;CA</td>
<td>150,000,000</td>
</tr>
</tbody>
</table>

A 7.2K/15k rpm HDD can only do 100/200 IOPS*

* (typical/worst cases considered)

For the time it takes to do Each Disk Operation:
- Millions of CPU Operations can be done
- Hundreds of Thousands of Memory Operations can be accomplished
Flash Memory Filling Price/Perf. Gaps

NVMe: The NextGen Interface for Solid State Storage

© 2012 Storage Networking Industry Association. All Rights Reserved.

Source: Flash Industry Report  IMEX Research ©2009-12

Best Opportunity to fill the gap is for storage to be close to Server CPU.

Source: IMEX Research SSD Industry Report ©2010-12
Data I/O Path – Insights

1. Latency and Processing Time
   - Milliseconds
   - Servers

2. Microseconds
   - CPUs
   - NAND Flash
   - or NVDIMMs
   - or SATA DIMMs

3. NVMes: The NextGen Interface for Solid State Storage
   - © 2012 Storage Networking Industry Association. All Rights Reserved.

Source: Flash Industry Report IMEX Research ©2009-12
New Storage Hierarchy in NGDC & Clouds

I/O Access Frequency vs. Percent of Corporate Data 2015

Cloud Storage
- Primary Capacity Storage
- Back Up Data
- Archived Data
- Offsite DataVault

SSD
- Logs
- Journals
- Temp Tables
- Hot Tables

FCoE/ SAS Arrays
- Tables
- Indices
- Hot Data
- Primary Storage

Source: Flash Industry Report IMEX Research ©2009-12

© 2012 Storage Networking Industry Association. All Rights Reserved.
Enterprise SSDs - Market Opportunity
By Interface

PCIe SSDs – Big Market Opportunity
NVM Express
A High performance, Scalable Host Controller Interface designed for Enterprise and Client systems using PCIe based SSDs.

Developed to reduce latency and provide faster performance with support for security and end-to-end data protection, the specification provides a flexible architecture with a streamlined register interface and command.

Defined by 80+ NVM Express Work Group members, it is supported by key industry leaders such as Dell, EMC, IDT, Micron, Intel, LSI, Microsoft, NetApp, Oracle, STEC, Cisco, Samsung, Marvell, SanDisk etc.

Expect a DRAM SSD to outperform today’s NAND Flash SSDs.
<table>
<thead>
<tr>
<th>Standard</th>
<th>Elements/Targets/Status</th>
<th>Pros/Cons</th>
</tr>
</thead>
</table>
| **PCI Express (PCIe)**                       | • PCIe 3.0 Shipping in High Volume  
  - HPC 4Q2011, Enterprise 1Q2012, Client 2Q2012  
  - 2.5”/3.5” FF Products Available  
  - Mutiple products Available                                                                                                                                 | • Low Latency, High Performance, Low Power targeted                                                 |
| **SSD Form Factor Working Group**            | • Multiple Protocols PCIe3.0/SAS 3.0/SATAExpress 3.0 (57 Members)  
  • Unified Connector (SFF8639) for PCIe 3.0, Multilane SAS, SATAe compatible with SAS/SATA                                                                 | • Hot Plug Connector  
  • Defined Required Syst Behaviour                                                                                                                            |
| **NVM Express (NVMe)**                       | • Spec 1.0 completed Mar 2011  
  • 80+ Companies participating  
  • Windows & Linux OS Drivers Available  
  • Multiple Products in Development                                                                                                                           | • NVM Storage I/F for Client Plus Server.  
  • Newer I/F vs. SCSI – can accommodates and/or by-pass SCSI stack                                     |
PCle Strengths

High performance
– Very Low Latency (no HBA overhead)
– Scalable link speed (250/500/1000 MB/s)
– Scalable port width (x1 to x16)
– Full duplex, multiple Queues requests, OOO processing

Low cost
– High volume standard commodity product
– Eliminates HBA cost

Power management capabilities
– Direct attach to CPU eliminates HBA power
– Features: Link power management, Dynamic Power Allocation, Optimized Buffer Flush/Fill
– Optimized link idle power with L1.OFF

Industry Standard
– Adopted by Several manufacturers
To reduce bottlenecks from legacy storage stacks, expect NVM Express to reduce latency overhead by greater than 50%
## PERFORMANCE - NVMe DRAM SSD vs PCIe SSD
*(Clocks per IO) Lower is Better*

<table>
<thead>
<tr>
<th>Environment/Workload</th>
<th>DRAM SSD NVM Express</th>
<th>PCIe SSD Fusion-io Drive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaming</td>
<td>12,986</td>
<td>1,5023</td>
</tr>
<tr>
<td>NASTRAN</td>
<td>8,049</td>
<td>12467</td>
</tr>
</tbody>
</table>

### System Configuration
- 2-Socket ES-2680 platform
- 32GB DDR3-1333 MHz RAM
- Red Hat Enterprise Linux
- SSDs Compared: Fusion-io 640 GB io Drive
- NVM Express Prototype PCIe SSD (DRAM Based)
## NVMe PCIe Flash Products - Form Factors

<table>
<thead>
<tr>
<th>Present PCIe Flash FF</th>
<th>New Standard SSD FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Board HHHL or FHHL</td>
<td>2.5” Standard SSD FF Proposed</td>
</tr>
</tbody>
</table>

Source: Flash Industry Report IMEX Resaerch ©2009-12
## NVMe PCIe Flash Products - Form Factors

<table>
<thead>
<tr>
<th>Present PCIe Flash FF</th>
<th>New Standard SSD FF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Board</strong> HHHL or FHHL</td>
<td><strong>2.5” Standard SSD FF Proposed</strong></td>
</tr>
<tr>
<td><strong>Shortcomings</strong></td>
<td><strong>Advantages</strong></td>
</tr>
<tr>
<td>- Form factor isn’t very rugged, with exposed components requiring careful handling</td>
<td>- Allows SSD to share a common backplane on Std. front end of the server (same as SATA and SAS drives)</td>
</tr>
<tr>
<td>- To install/swap these cards in a traditional server, have to power down the server i.e. take it out of commission &gt;</td>
<td>- Scalability – Adding more density on demand</td>
</tr>
<tr>
<td>- Big problem in cloud or database environment where servers are in full 100%, Always-On utilization</td>
<td>- Serviceability - Rugged form factor, no need to power</td>
</tr>
<tr>
<td></td>
<td>- Improved efficiency – Higher performance</td>
</tr>
<tr>
<td></td>
<td>- Lower TCO – Lower cost/IOPs</td>
</tr>
</tbody>
</table>
NVMe: A Std. Interoperable Interface

- Defined by 80+ NVM Express Work Group members,
- Supported by major industry players

DSSD, Enmotus, Eonsil, ExpertIO, Fresco Logic, Fujitsu, GenesysLogic, Hitachi GST, Huawei, Symantec, Hyperstone, IP Cores,

LeCroy, Link-A-Media Devices, LSI SandForce, Mobile Semiconductor, NVELO, OCZ Technology, Phison Electronics,

PLX Technology, Qlogic, Silicon Motion, Toshiba, ULINK Technology, VIA, Violin Memory, Virident, Western Digital, Xyratex (as of Sep 2012)
### Data Center Performance Bottlenecks

<table>
<thead>
<tr>
<th></th>
<th>SATA</th>
<th>SAS</th>
<th>PCIe</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DriveForm Factors</strong></td>
<td>SATA</td>
<td>SAS</td>
<td>Multilink</td>
<td>SOP/PQI</td>
</tr>
<tr>
<td>1.8”, 2.5”, 3.5”</td>
<td>2.5”, 3.5”</td>
<td>2.5”</td>
<td>2.5”</td>
<td>2.5”, Card</td>
</tr>
<tr>
<td><strong>No of Ports/ Lanes</strong></td>
<td>1</td>
<td>1, 2</td>
<td>1, 2, 4</td>
<td>1, 2, 4</td>
</tr>
<tr>
<td><strong>Command Set/Que Interface</strong></td>
<td>ATA / SATA-IO</td>
<td>SCSI / SAS</td>
<td>SCSI / SAS</td>
<td>SCSI / SOP/PQI</td>
</tr>
<tr>
<td><strong>Transfer Rate</strong></td>
<td>6Gb/s</td>
<td>12Gb/s</td>
<td>12Gb/s</td>
<td>8 Gb/s</td>
</tr>
<tr>
<td><strong>DriveConnector</strong></td>
<td>SFF-xxxxx</td>
<td>SFF-8680</td>
<td>SFF-8639</td>
<td>SFF-8639</td>
</tr>
<tr>
<td><strong>Express Bay Compatible?</strong></td>
<td>Yes, 2.5”</td>
<td>Yes, 2.5”</td>
<td>Yes, 2.5”</td>
<td>Yes, 2.5”</td>
</tr>
<tr>
<td><strong>Drive Power (Typical)</strong></td>
<td>9W Typical</td>
<td>9W Typical</td>
<td>Upto 25W</td>
<td>Upto 25W</td>
</tr>
<tr>
<td><strong>Max Bandwidth</strong></td>
<td>0.6GB/s (x2)</td>
<td>4.8 GB/s (x4)</td>
<td>9.6GB/s (x4)</td>
<td>8 GB/s (x4)</td>
</tr>
<tr>
<td><strong>Host DriverStack (Stg Cntlr/Direct Drives)</strong></td>
<td>AHCI</td>
<td>IHV</td>
<td>IHV</td>
<td>Common Driver (SOP/PQI)</td>
</tr>
</tbody>
</table>

**Source:** Flash Industry Report IMEX Research ©2009-12

© 2012 Storage Networking Industry Association. All Rights Reserved.
NVMe Usage Models

- Used for temporary data
- Non-redundant
- Used to reduce memory footprint

- Typically for persistent data
- Redundant (i.e., RAID’ed)
- Commonly used as Tier-0 storage

- Used for Metadata or data
- Multi-ported device
- Redundancy based on usage
Key Takeaways

- **Solid State Storage creating a paradigm shift in Storage Industry**
  - Leverage the opportunity to optimize your computing infrastructure with SSD adoption after making a due diligence in selection of vendors/products, industry testing and interoperability

- **Goal: Optimize Transactions for Query Response Time vs. # of Users**
  - Improving Query Response time for a given number of users (IOPs) or Serving more users (IOPS) for a given query response time

- **Server Based Flash Storage outperforms Storage Based Flash**
  - Reduce the latency gap between CPUs and Flash NAND Storage by placing NAND Flash in Servers and avoiding latency caused by HBAs, Switches, Networks, Storage Controllers.

- **Utilize NVMe Interface for PCIe based Flash**
  - Achieve a 50% improvement using NVM Express over using a SCSI?SAS storage stack.
The SNIA Education Committee would like to thank the following individuals for their contributions to this Tutorial.

Authorship History

Anil Vasudeva, April 2012

Updates:
Anil Vasudeva, October 2012

Additional Contributors

Late Jim Gray (Pioneer in Data I/O Insights)

Please send any questions or comments regarding this SNIA Tutorial to tracktutorials@snia.org