The Performance Impact of NVMe and NVMe over Fabrics

Live: November 13, 2014

Presented by experts from
Cisco, EMC and Intel

Webcast Presenters

SNIA
Ethernet Storage Forum

- J Metz, R&D Engineer for the Office of the CTO, Cisco
- David Akerson, (need title), Intel
- Amber Huffman, Senior Principal Engineer, Intel
- Steve Sardella , (need title), EMC
- Dave Minturn, (need title), Intel

SNIA Legal Notice

SNIA Ethernet Storage Forum

- The material contained in this tutorial is copyrighted by the SNIA 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.

What This Presentation *Is*

SNIA
Ethernet Storage Forum

- A discussion of a new way of talking to Non-Volatile Memory (NVM)
- Technical information about how it works and what it's for
- Exploration of possible uses and methods to implement
- Identification of related technologies

What This Presentation Is *Not*

SNIA
Ethernet Storage Forum

- The final say!
- Discussion of products or vendors
- Recommendation for uses

Agenda

SNIA
Ethernet Storage Forum

- NVM Express Genesis
- NVM Express: Transforming IT Infrastructures
- Extending NVM Express Efficiency: NVMe over Fabrics
- Expert Round Table

What's the Difference between NVM and NVMe?

SNIA Ethernet Storage Forum

- **NVM** stands for ***Non-Volatile Memory***
  - Flash, SSDs, NVDIMMs, all qualify
- **NVMe** stands for ***NVM Express***
  - An interface to the controller for NVM
  - A mechanism for providing commands to the drives

Examples of NVM

NVMe is Architected for NVM

SNIA Ethernet Storage Forum

- NVM Express is a standardized high performance software interface for PCIe SSDs
  - Standardizes register set, feature set, and command set where there were only proprietary PCIe solutions before
  - Architected from the ground up for NAND and next generation NVM
  - Designed to scale from Enterprise to Client systems

- Developed by an open industry consortium with a 13 company Promoter Group

PCI Express SSD Benefits

SNIA Ethernet Storage Forum

Lower latency: Direct connection to CPU

Scalable performance: 1 GB/s per lane – 4 GB/s, 8 GB/s, … in one SSD

Industry standards: NVM Express and PCI Express (PCIe) 3.0

Increased I/O: Up to 40 PCIe lanes per CPU socket

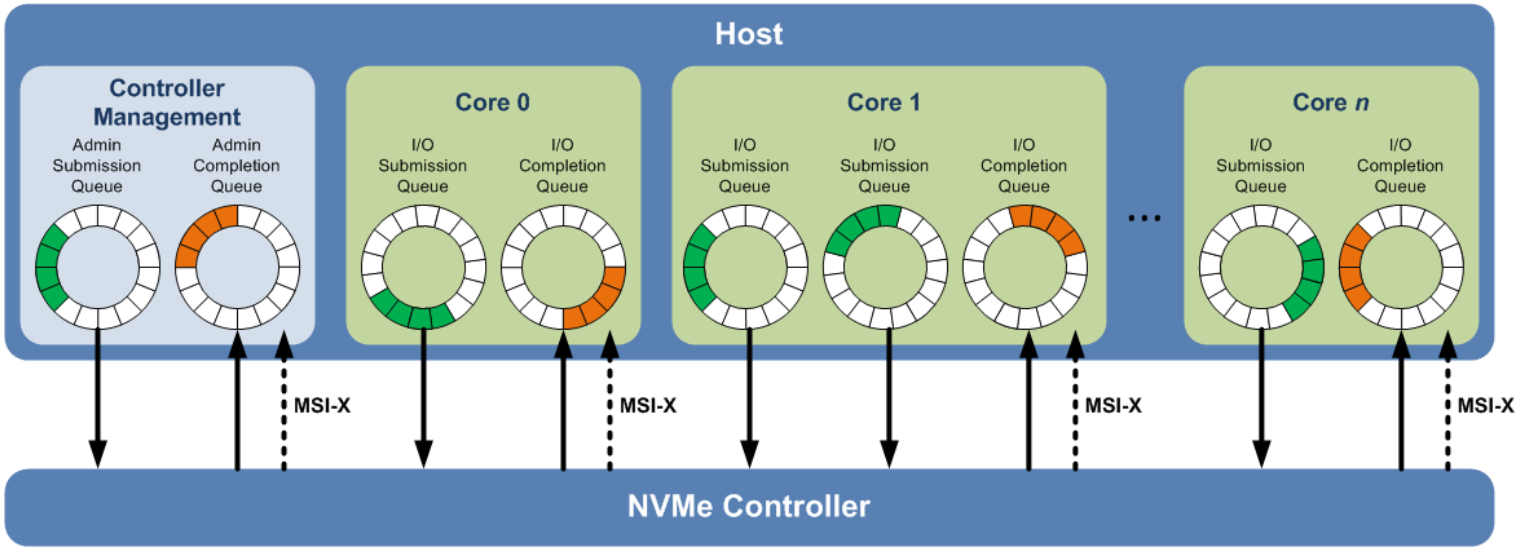
Security protocols: Trusted Computing Group Opal

Low Power features: Low power link (L1.2), NVMe power states

Form factors: SFF-8639, SATA Express*, M.2, Add in card, Future: BGA (PCI SIG)

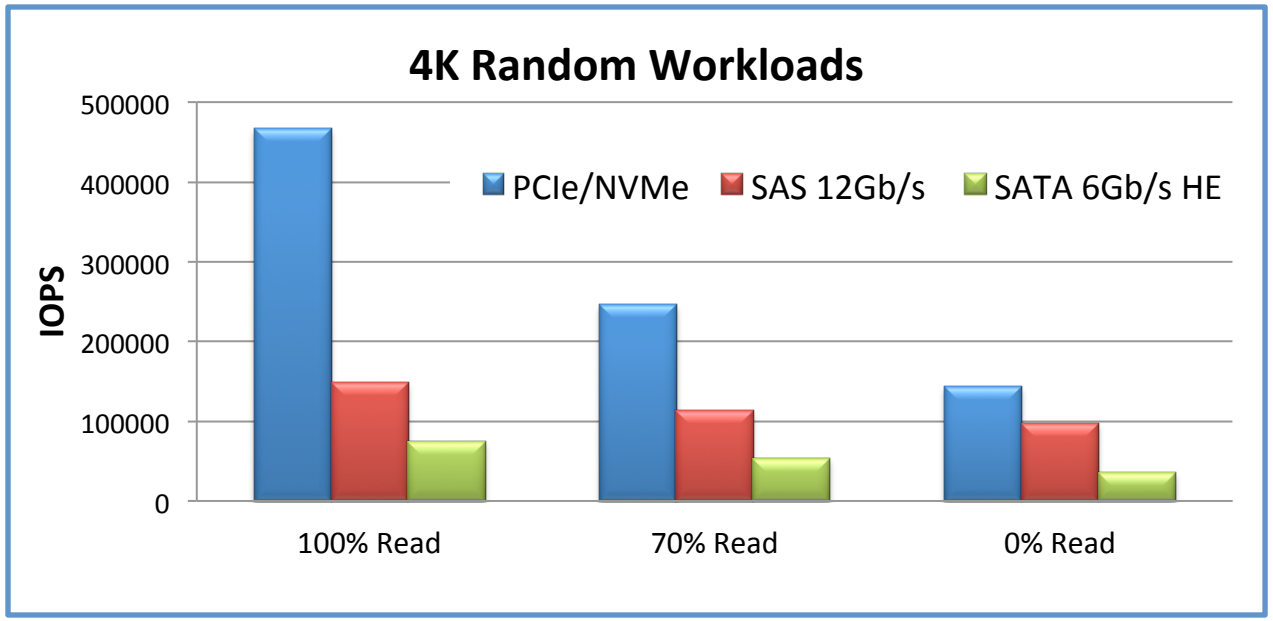
NVMe Technical Overview

- Supports deep queues (64K commands per queue, up to 64K queues)
- Supports MSI-X and interrupt steering
- Streamlined & simple command set (13 required commands)
- Optional features to address target segment
  - Data Center: End-to-end data protection, reservations, etc.
  - Client: Autonomous power state transitions, etc.
- Designed to scale for next generation NVM, agnostic to NVM type used



NVMe Delivers Best in Class IOPs

- 100% random reads: NVMe has > 3X better IOPs than SAS 12Gbps
- 70% random reads:NVMe has > 2X better IOPs than SAS 12Gbps
- 100% random writes: NVMe has ~ 1.5X better IOPs than SAS 12Gbps

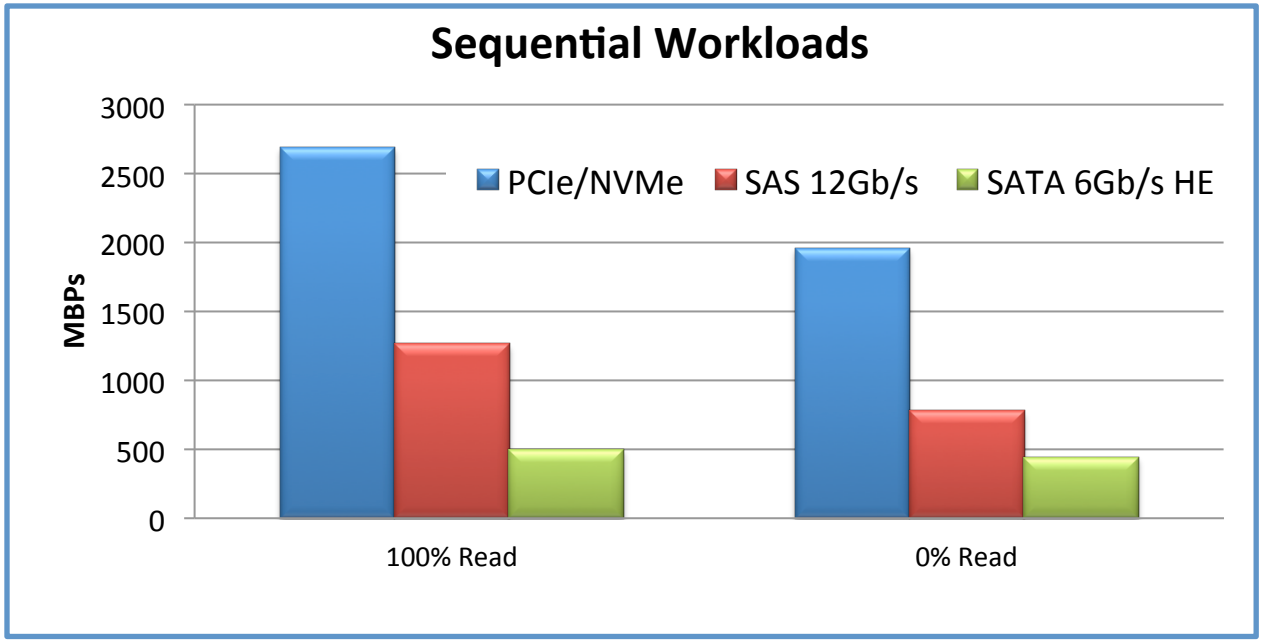


CI Express* (PCIe*)/NVM Express* (NVMe) Measurements made on Intel® Core™ i7-3770S system @ 3.1GHz and 4GB Mem running Windows* Server 2012 Standard O/S, Intel PCIe/NVMe SSDs, data collected by IOmeter* tool. PCIe/NVMe SSD is under ent. SAS Measurements from HGST Ultrastar* SSD800M/1000M (SAS) Solid State Drive Specification. SATA Measurements from Intel Solid State Drive DC P3700 Series Product Specification. Software and workloads used in performance tests may have imized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Best in Class Sequential Performance

SNIA Ethernet Storage Forum

- NVM Express (NVMe) delivers > 2.5GB/s of read and ~ 2 GB/s of write performance
  - 100% reads: NVMe has >2X better performance than SAS 12Gbps
  - 100% writes: NVMe has >2.5X better performance than SAS 12Gbps



Note: PCI Express* (PCIe*)/NVMe Measurements made on Intel® Core™ i7-3770S system @ 3.1GHz and 4GB Mem running Windows* Server 2012 Standard O/S, Intel PCIe/NVMe SSDs, data collected by IOmeter* tool. PCIe/NVMe SSD is under development. SAS Measurements from HGST Ultrastar* SSD800M/1000M (SAS) Solid State Drive Specification. SATA Measurements from Intel Solid State Drive DC P3700 Series Product Specification. Source: Intel Internal Testing. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Analyzing What Matters

SNIA Ethernet Storage Forum

- What matters in today's Data Center is not just IOPs and bandwidth
- Let's look at efficiency of the software stack, latency, and consistency

Server Setup

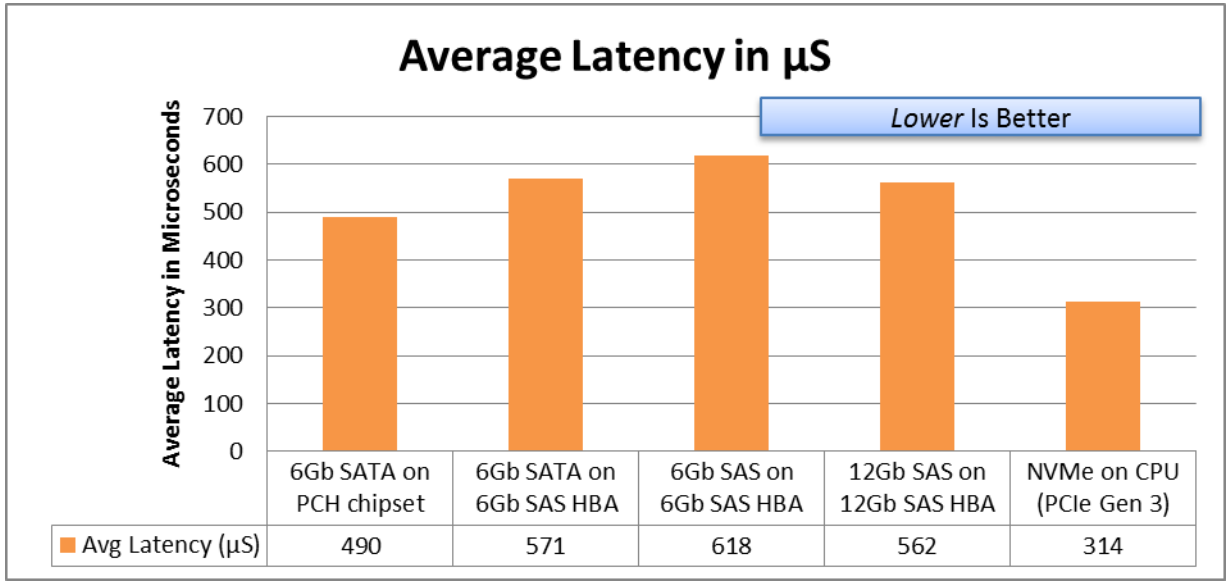
- Basic 4U Intel® Xeon® E5 processor based server
- Out of box software setup
- Moderate workload: 8 workers, QD=4, random reads

Storage Protocols Evaluated					
Interface	6Gb SATA	6Gb SATA	6Gb SAS	12Gb SAS	NVMe PCIe Gen 3
Attach Point	PCH chipset	6Gb SAS HBA	6Gb SAS HBA	12Gb SAS HBA	CPU

Not strenuous on purpose – evaluate protocol and not the server.

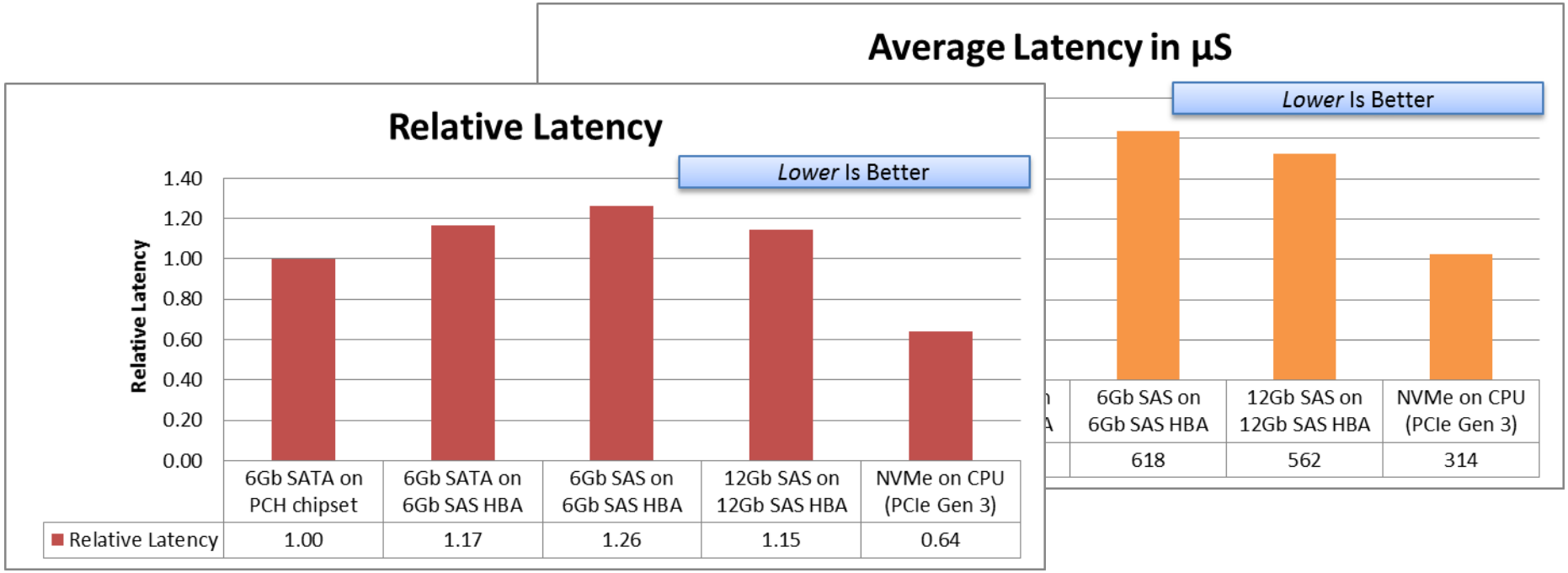
Latency of NVM Express

- The efficiency of NVMe directly results in leadership latency
- NVMe is more than 200 µs lower latency than 12 Gb SAS



Latency of NVM Express

- The efficiency of NVMe directly results in leadership latency
- NVMe is more than 200 µs lower latency than 12 Gb SAS



NVMe delivers the lowest latency of any standard storage interface.

Agenda

SNIA
Ethernet Storage Forum

- NVM Express Genesis
- NVM Express: Transforming IT Infrastructures
- Extending NVM Express Efficiency: NVMe over Fabrics
- Expert Round Table

EMC's Perspective: A Look Back

- Several years ago, the storage industry was at a crossroad, with regard to SSDs
- SAS and SATA SSDs were popular, but it was expected that use of PCI Express SSDs would grow dramatically, due to performance benefits
- Without standardization, there would have been many disparate hardware and software solutions for PCIe SSDs
- A group of companies joined together and created the NVM Express Specification

"EMC is the leader in providing SSDs to the market in enterprise storage systems, and believes standards are in the best interest of the industry"

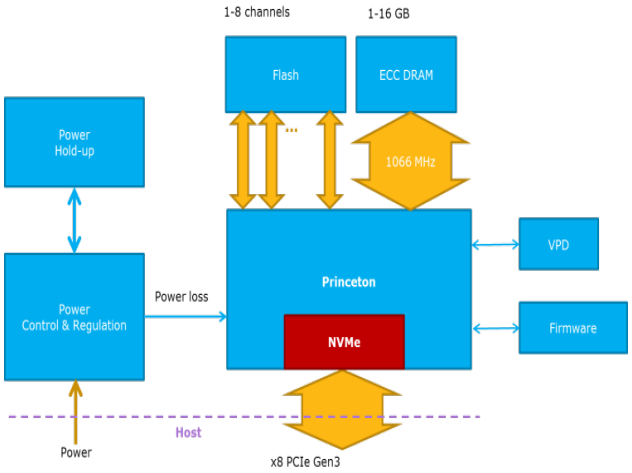
"As a driving force in enterprise Flash technology innovation, EMC recognizes the importance of NVMe to expanding the adoption of Flash."

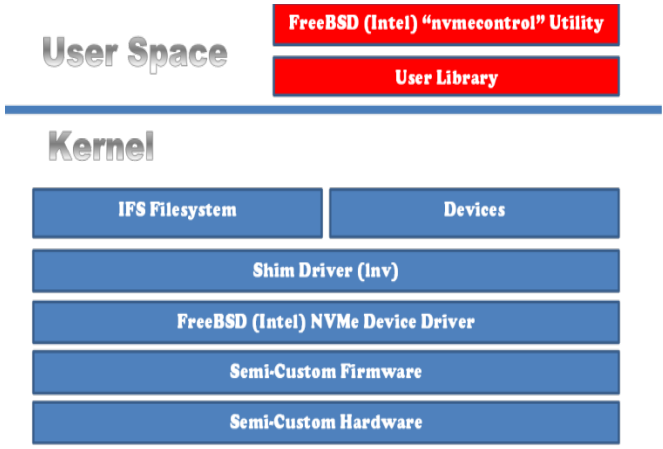
"By standardizing many enterprise-class features, the NVMe specification is enabling the industry to deliver higher performance PCIe SSDs without compromising on expectations for drive functionality—which is especially important as customers continue transforming their IT infrastructures."

Bill DePatie
Senior Vice President, Global Platform Engineering
EMC

EMC's Perspective: The Present

- Flash has become an indispensable component of Data Center storage
  - EMC, through development and acquisition, has invested in Flash at every level, including:
    - Server Flash
    - Hybrid Arrays
    - All-Flash Arrays
- Future Non-Volatile Media holds the promise of even greater performance and capabilities
  - The NVMe specification defines a highly efficient standardized interface for the media of today and tomorrow
- EMC was able to reduce the hardware/software development and validation time for a high performance NVRAM design, thanks to NVM Express controllers and drivers
  - More info can be found at:





EMC's Perspective: A Look Forward

SNIA Ethernet Storage Forum

- Once again, the storage industry is at a crossroad
  - This time, with respect to "NVMe over X", where "X" could be any existing I/O protocol
- The NVMe Specification did such a good job of defining an efficient queuing interface for storage, there is now a desire to extend it to other protocols
  - These mature protocols are already established within the Data Center, and have certain advantages over PCI Express, in terms of robustness and error handling
- Without standardization, there could be many disparate implementations, by protocol or by silicon vendor
- The NVMe group has taken up the call to address this, with "NVMe over Fabrics"

"EMC is pleased to be a core contributor to the definition of NVMe over Fabrics, the new NVM Express standard for sharing next-generation flash storage in an RDMA-capable fabric. " Mike Shapiro, Vice President, Software Engineering of DSSD EMC

Agenda

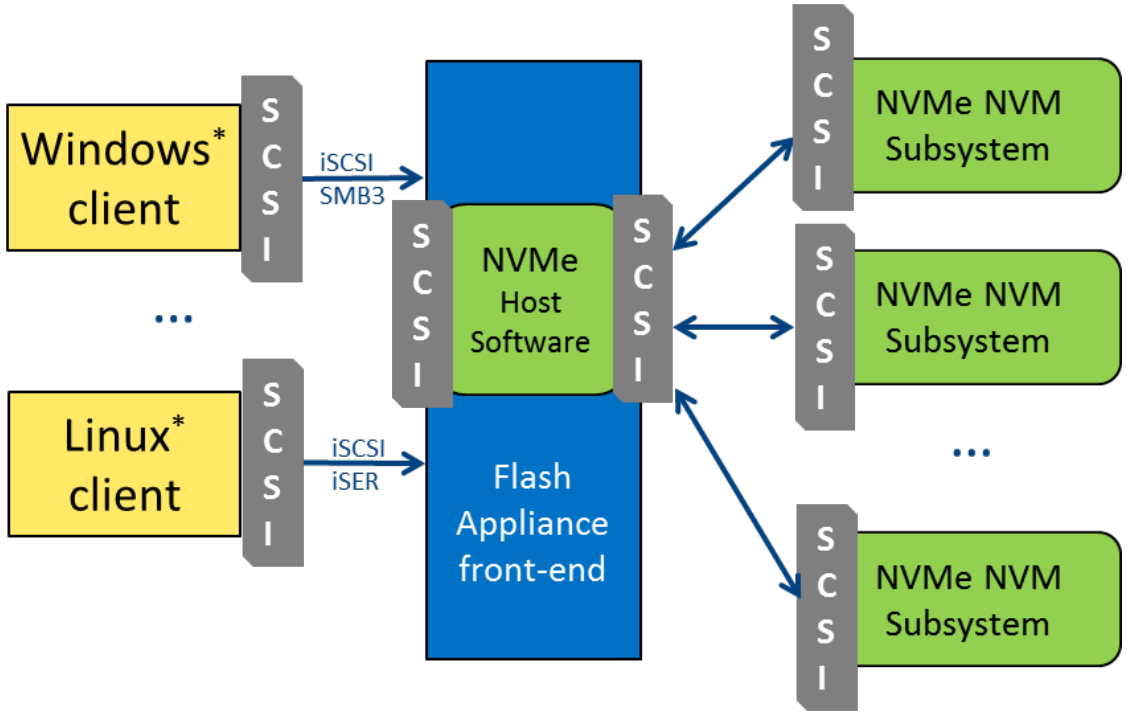
SNIA
Ethernet Storage Forum

- NVM Express Genesis
- NVM Express: Transforming IT Infrastructures
- Extending NVM Express Efficiency: NVMe over Fabrics
- Expert Round Table

NVM Express (NVMe) in Non-PCI Express Fabric Environments

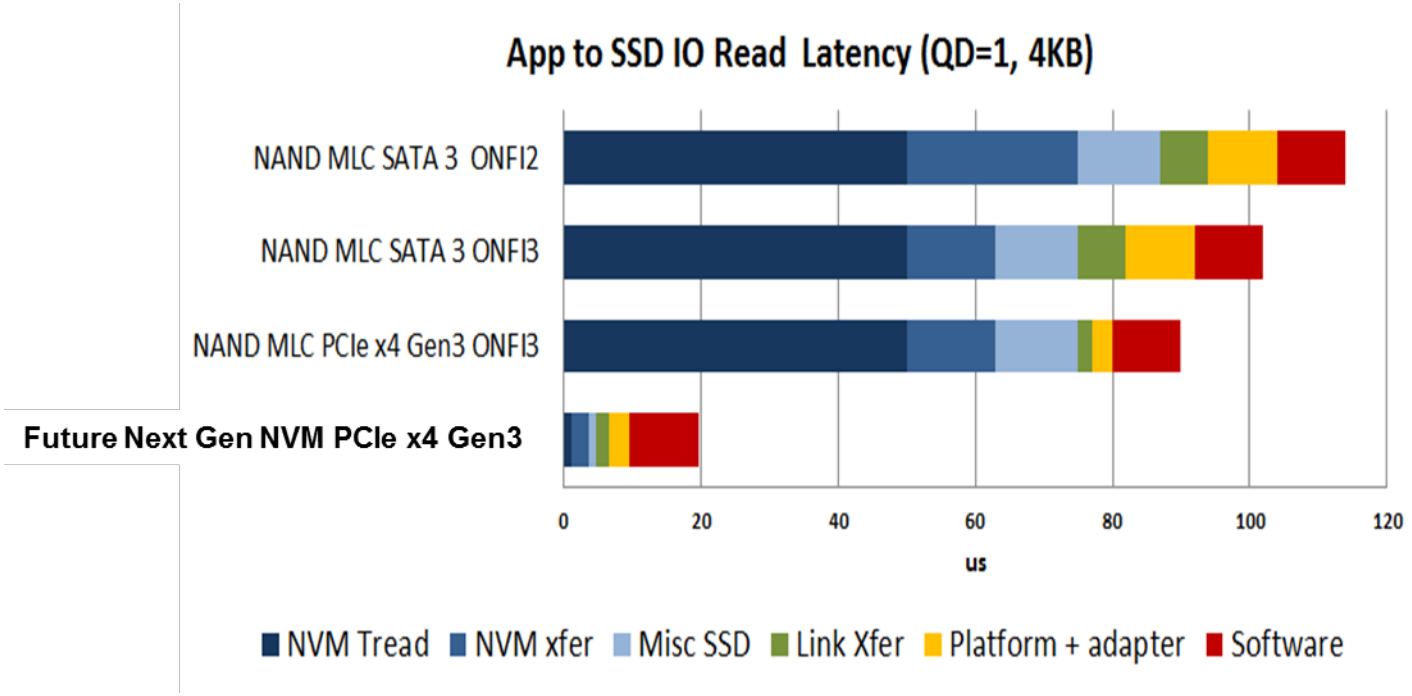
SNIA Ethernet Storage Forum

- A primary use case for NVM Express (NVMe) is in a Flash Appliance
- Hundreds or more SSDs may be attached – too many for PCI Express based attach
- Concern: Remote SSD attach over a fabric uses SCSI based protocols today – requiring protocol translation(s)



Desire best performance and latency from SSD investment over fabrics like Ethernet, InfiniBand™, Fibre Channel, and Intel® Omni Scale Fabric.

Realizing Benefit of Next Gen NVM over Fabrics



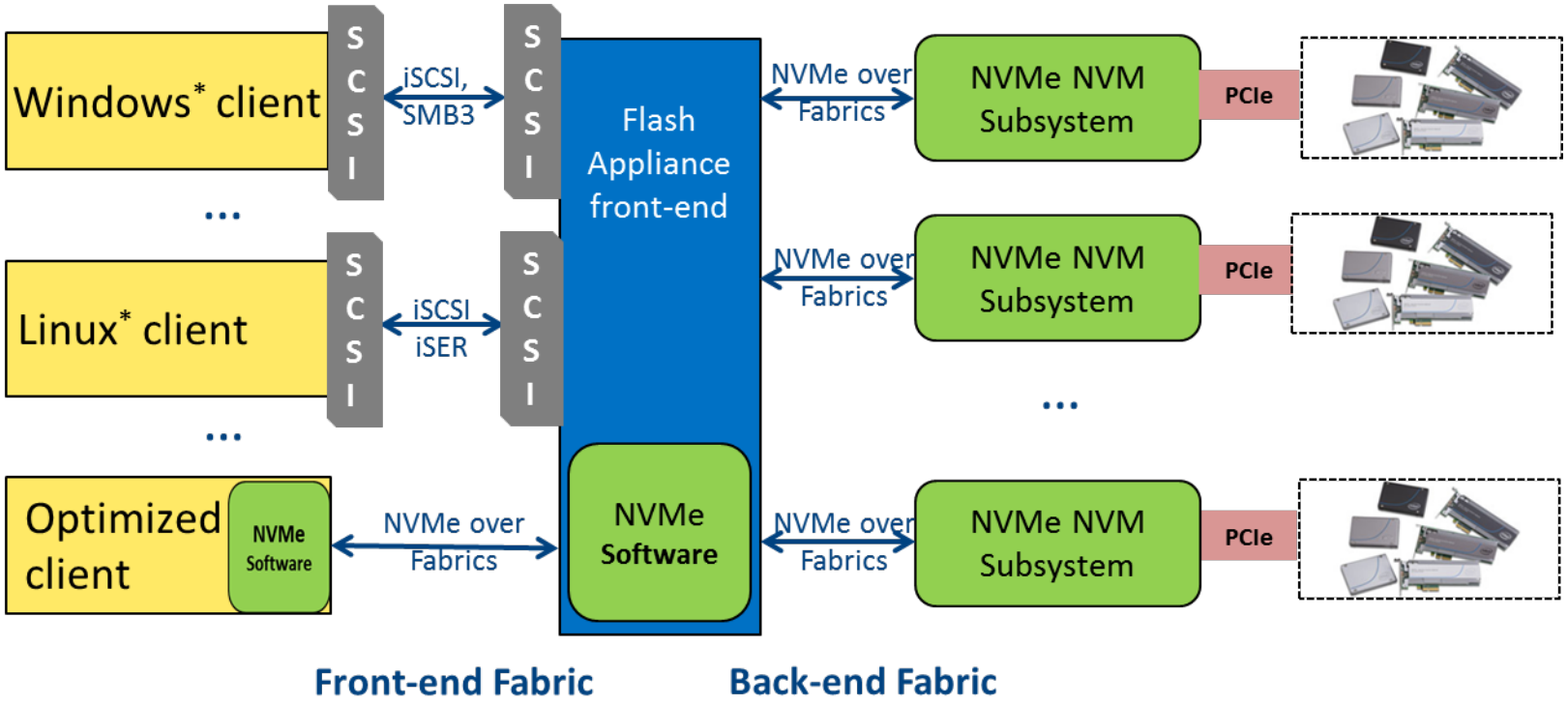
- NVM Express (NVMe) SSD latency may be < 10 µs with next generation NVM
- Using a SCSI-based protocol for remote NVMe adds over 100 µs in latency

Concern: Low latency of next gen NVM lost in (SCSI) translation.

Source: Intel Measurements.

Introducing NVM Express (NVMe) over Fabrics

SNIA Ethernet Storage Forum



Extend efficiency of NVMe over front and back-end fabrics.

Why NVM Express over Fabrics?

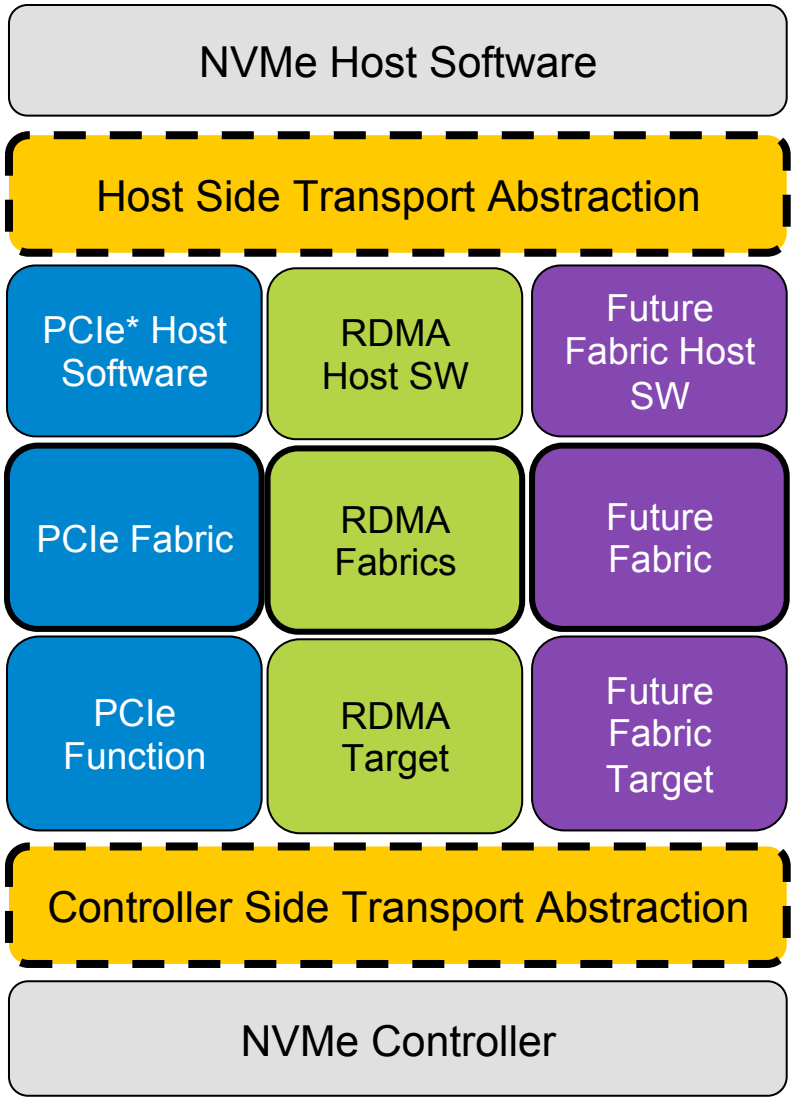
SNIA Ethernet Storage Forum

- Simplicity, Efficiency and End-to-End NVM Express (NVMe) Model
  - NVMe supports up to 64K I/O Queues with 3 required commands
  - Inherent parallelism of multiple I/O Queues is exposed
  - Simplicity of protocol enables hardware automated I/O Queues – transport bridge
  - No translation to or from another protocol like SCSI (in firmware/software)
  - NVMe commands and structures are transferred end-to-end
  - Maintains consistency between fabric types by standardizing a common abstraction

Goal: Make remote NVMe equivalent to local NVMe, within ~ 10 µs latency.

Architectural Approach

- The NVM Express (NVMe) Workgroup has started the definition of NVMe over Fabrics
- A flexible transport abstraction layer is under definition, enabling a consistent definition of NVMe over many different fabrics types
- The first fabric definition is the RDMA protocol family – used with Ethernet (iWARP and RoCE) and InfiniBand™
- Expect future fabric definitions; such as Fibre Channel and Intel® Omni-Scale fabrics

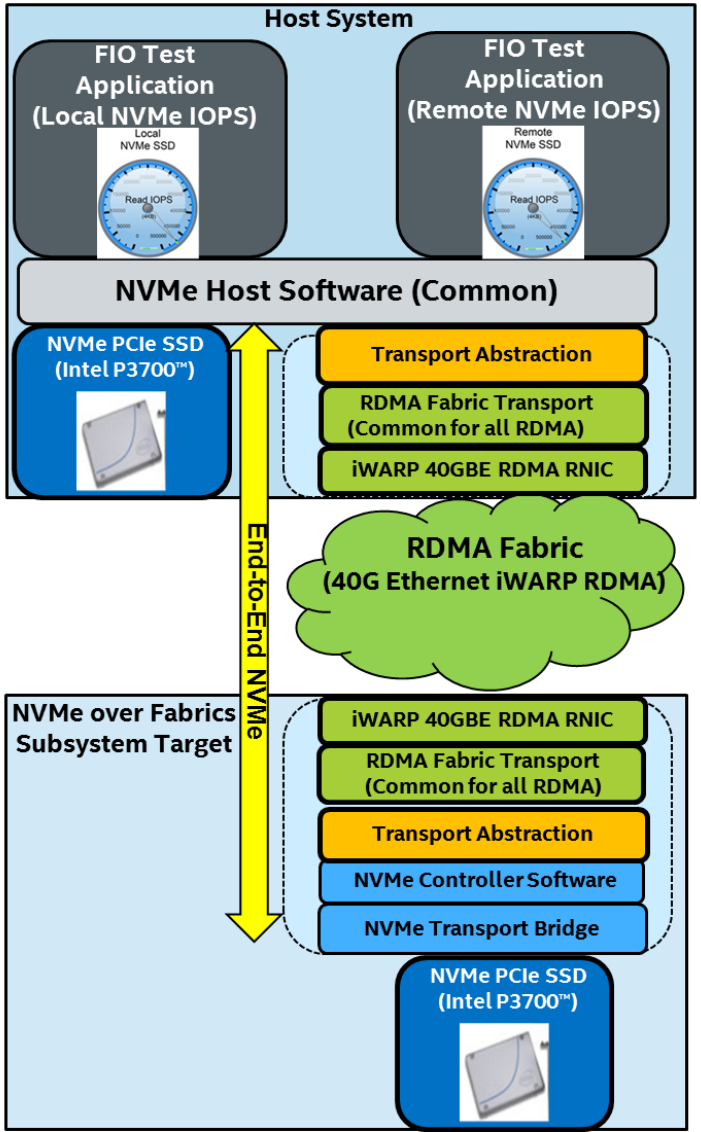


NVMe over Fabrics Prototype on iWARP

SNIA Ethernet Storage Forum

- Recall: Goal is remote NVM Express (NVMe) equivalent IOPS to local NVMe and no more than 10 µs added latency
- Prototype delivers 460K IOPs for ***both*** the local and remote PCIe NVMe SSD devices
- Remote NVMe adds ***8 µs*** latency versus local NVMe access (4K Read & Write; QD=1)
- Demonstrates the efficiency of NVMe End-to-End; NVMe Target software running on one CPU core (two SMT threads) at 20% utilization

Get involved with NVMe over Fabrics definition.



Intel i7-4790 3.6GHz Processors, 8GB DDR-1600, Gibabyte GA-Z97X-UD7 MB, Intel P3700 800G SSDs, Chelsio T580-CR 40GBE iWARP NIC. RHEL7 Linux, OFED 3.2 Software, FIO V2.1.10. Source: Intel. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Agenda

SNIA
Ethernet Storage Forum

- NVM Express Genesis
- NVM Express: Transforming IT Infrastructures
- Extending NVM Express Efficiency: NVMe over Fabrics
- Expert Round Table

Expert Round Table

SNIA
Ethernet Storage Forum

- Discussion and Q&A

Summary

SNIA
Ethernet Storage Forum

- NVM Express (NVMe) is a great Data Center investment, near term and long term
- NVMe delivers the lowest latency of any standard storage interface
- Innovation continues – get involved in NVMe over Fabrics!

For more information, visit www.nvmexpress.org

After This Webcast

SNIA Ethernet Storage Forum

- This webcast will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
  - http://www.snia.org/forums/esf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog
  - http://sniaesfblog.org/
- Follow and contribute to the SNIA-ESF blog thread on many storage-over-Ethernet topics, both hardware and protocols
  - http://sniaesfblog.org/

Setup for Efficiency and Latency Analysis

SNIA Ethernet Storage Forum

- Server setup:
  - 2-Socket Intel® Xeon® E5-2690v2 + 64GB RAM + SSD Boot/Swap – EPSD 4U S2600CP Family
  - Linux* 2.6.32-461.el6.bz1091088.2.x86_64 #1 SMP Thu May 1 17:05:30 EDT 2014 x86_64 x86_64 x86_64 GNU/Linux
  - CentOS* 6.5 fresh build, yum –y update (no special kernel or driver)
- SSDs used:
  - LSI* 9207-8i + 6Gb SAS HGST* Drive @ 400GB & LSI 9207-8i + 6Gb SATA Intel® SSD DC S3700 @ 400GB
  - LSI 9300-8i + 12Gb SAS HGST Drive @ 400GB
  - Onboard SATA Controller + SATA Intel® SSD DC S3700 @ 400GB
  - Intel® SSD DC P3700 Series NVM Express* (NVMe) drive at 400GB
- FIO workload
  - fio --ioengine=libaio --description=100Read100Random --iodepth=4 --rw=randread --blocksize=4096 --size=100% --runtime=600 --time_based --numjobs=1 --name=/dev/nvme0n1 --name=/dev/nvme0n1 --name=/dev/nvme0n1 --name=/dev/nvme0n1 --name=/dev/nvme0n1 --name=/dev/nvme0n1 --name=/dev/nvme0n1 --name=/dev/nvme0n1 2>&1 | tee -a NVMeONpciE.log
  - 8x workers, QD4, random read, 4k block, 100% span of target, unformatted partition