OPENFABRICS ALLIANCE

2013 OFA Developer Workshop

NVM Express Maturing Fast

NVM Express Overview

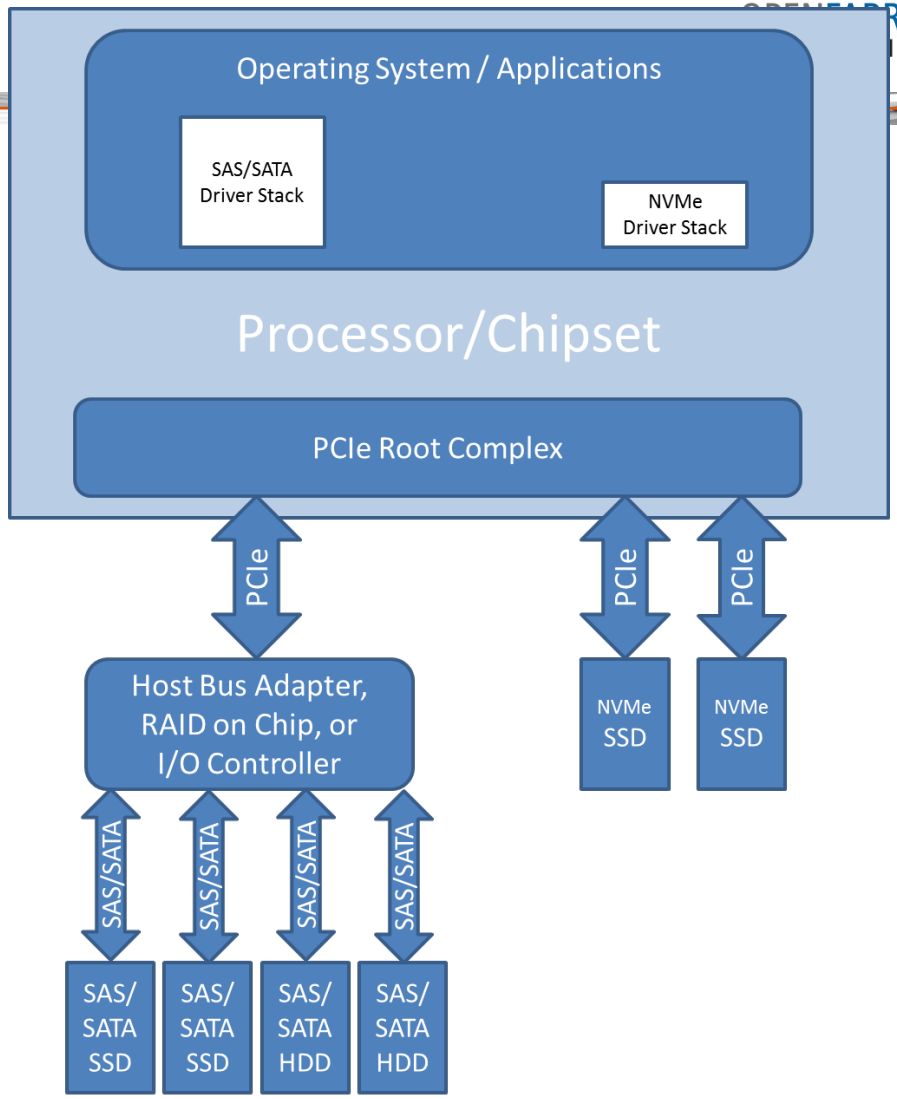
- NVM Express is a high performance, scalable host controller interface designed for Enterprise and client systems that use PCI Express* SSDs
- NVM technology agnostic
- NVMe developed by industry consortium of 80+ members and is directed by a 13-company Promoter Group

- Additional information at NVMExpress.org website

*Other names and brands may be claimed as the property of others.

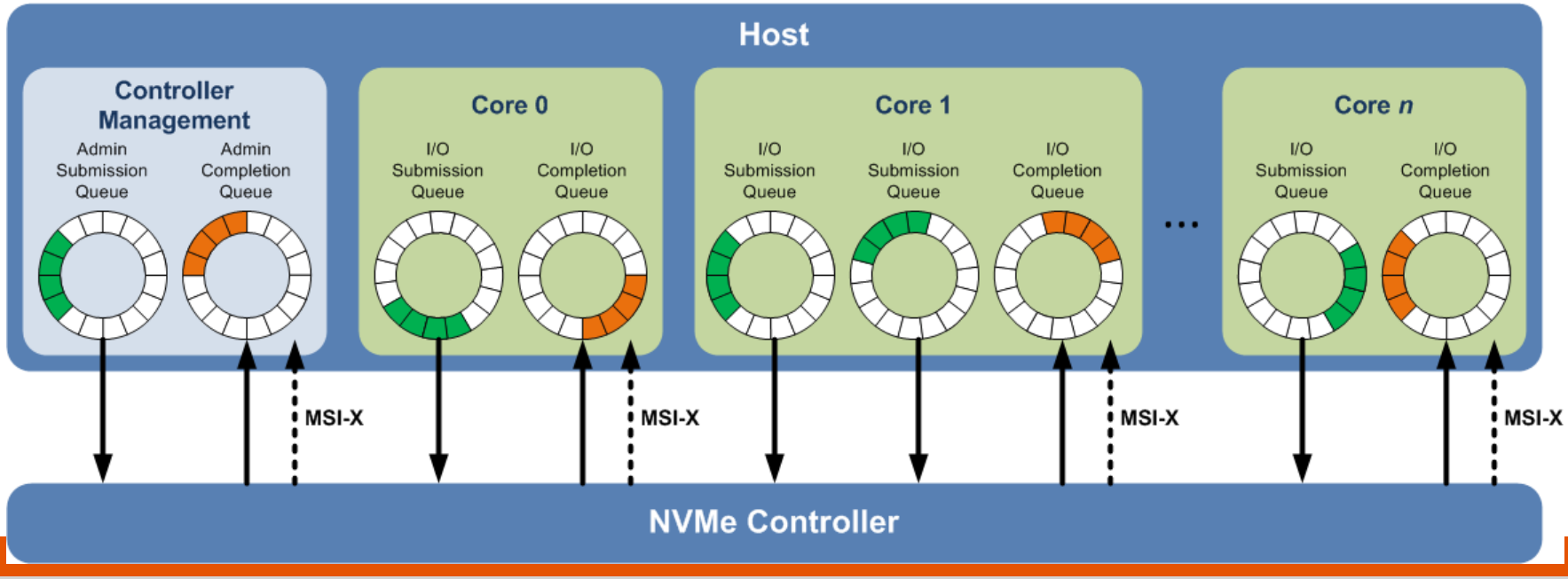
Architected for SSDs

- Traditional interfaces developed for HDD
  - Up to 200 IOPs
  - Inefficiencies hidden
- NVM Express architected for SSD
  - Over 3 million IOPs demo'd
  - Inefficiencies exposed



Technical Basics

- The focus of the effort is efficiency, scalability and performance
  - All parameters for 4KB command in single 64B DMA fetch
  - Supports deep queues (64K commands per Q, up to 64K queues)
  - Supports MSI-X and interrupt steering
  - Streamlined command set optimized for NVM (6 I/O commands)
  - Enterprise: Support for end-to-end data protection (i.e., DIF/DIX)



Streamlined Command Set

OPENFABRICS ALLIANCE

Management Commands for Queues & Transport

Admin Command	Description
Create I/O Submission Queue	Queue Management
Create I/O Completion Queue	
Delete I/O Submission Queue	
Delete I/O Completion Queue	
Abort	
Asynchronous Event Request	Status & Event Reporting
Get Log Page	
Identify	Configuration
Set Features	
Get Features	
(Optional) Firmware Activate	Firmware Management
(Optional) Firmware Image Download	
(Optional) Security Send	Security
(Optional) Security Receive	
(Optional) Format NVM	Namespace Management

I/O Commands for SSD Functionality

NVM Command	Description
Flush	Data Ordering
Read	Data Transfer, Including end-to-end data protection & security
Write	
(Optional) Write Uncorrectable	
(Optional) Compare	
(Optional) Dataset Management	Data Usage Hints

13 Required Commands Total (10 Admin, 3 I/O)

nvm EXPRESS

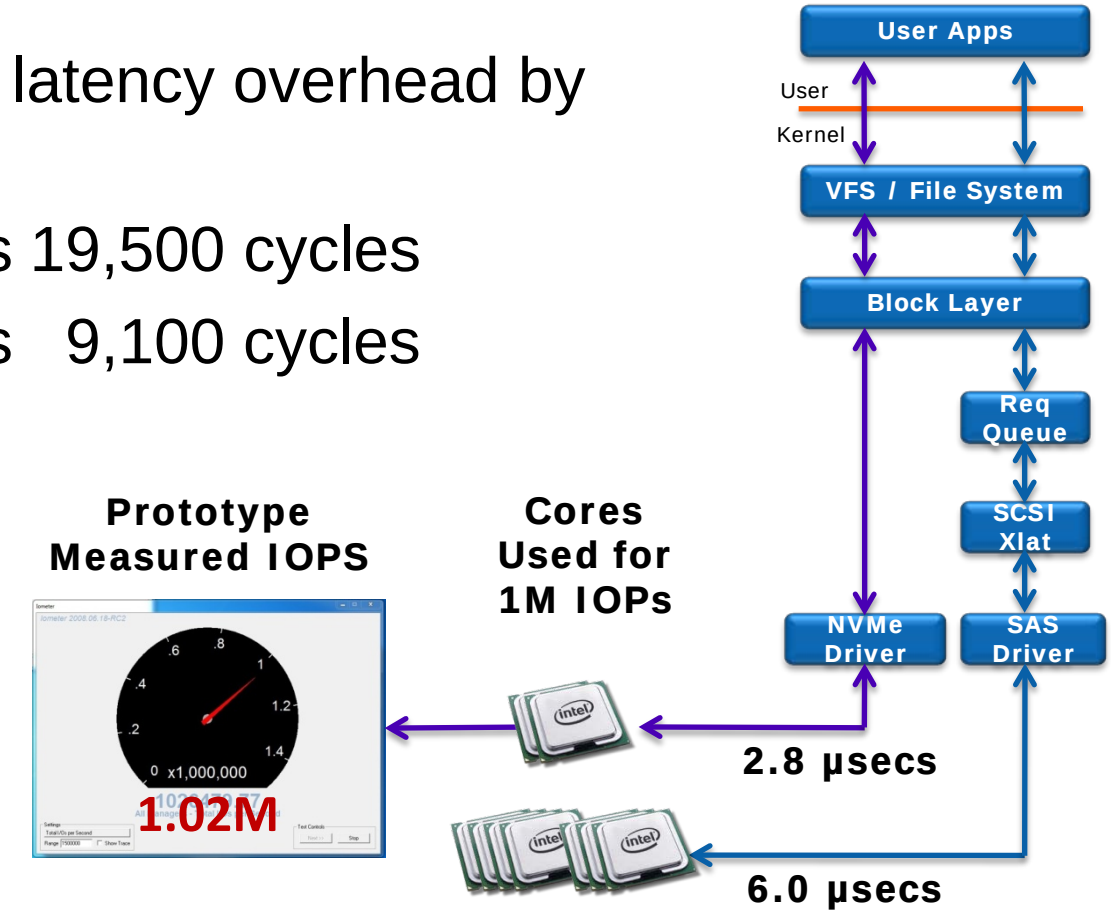
Case Study

NVM Express reduces latency overhead by **more than 50%**

SCSI/SAS: 6.0 µs 19,500 cycles

NVMe: 2.8 µs 9,100 cycles

Linux* Storage Stack



Chatham NVMe Prototype

Prototype Measured IOPS

Cores Used for 1M IOPs

Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux RedHat* EL6.0 2.6.32-71 Kernel.

NVM Express Status

- NVMe 1.0 published March, 2011
- NVMe 1.1 published October, 2012 adding Enterprise and Client capabilities
  - Enterprise: Multi-path I/O and namespace sharing
  - Client: Lower power through autonomous transitions during idle
- UNH-IOL NVMe plugfest scheduled for May, 13 to enable an interoperable ecosystem
- 1st products appearing in 2013, expect Enterprise / Data Center to adopt first

Reference Drivers

- Linux
  - Accepted into the mainline kernel on kernel.org
  - Open source GPL license
- Windows
  - Baseline developed in collaboration by IDT, Intel, and LSI
  - Open source BSD license
- VMware
  - Initial driver developed by Intel
  - “vmk linux” driver based on Linux version
- UEFI
  - Under development
  - Open source plan in 1H 2013
- Solaris
  - Working driver prototype

Summary

- NVM Express is optimized for PCIe SSDs, replacing standards designed for the HDDs
- Performance for today and tomorrow's NVM
- Product and driver infrastructure maturing quickly
- First NVMe PCIe SSDs expected in 2013

Additional information at NVMExpress.org website

Thank You

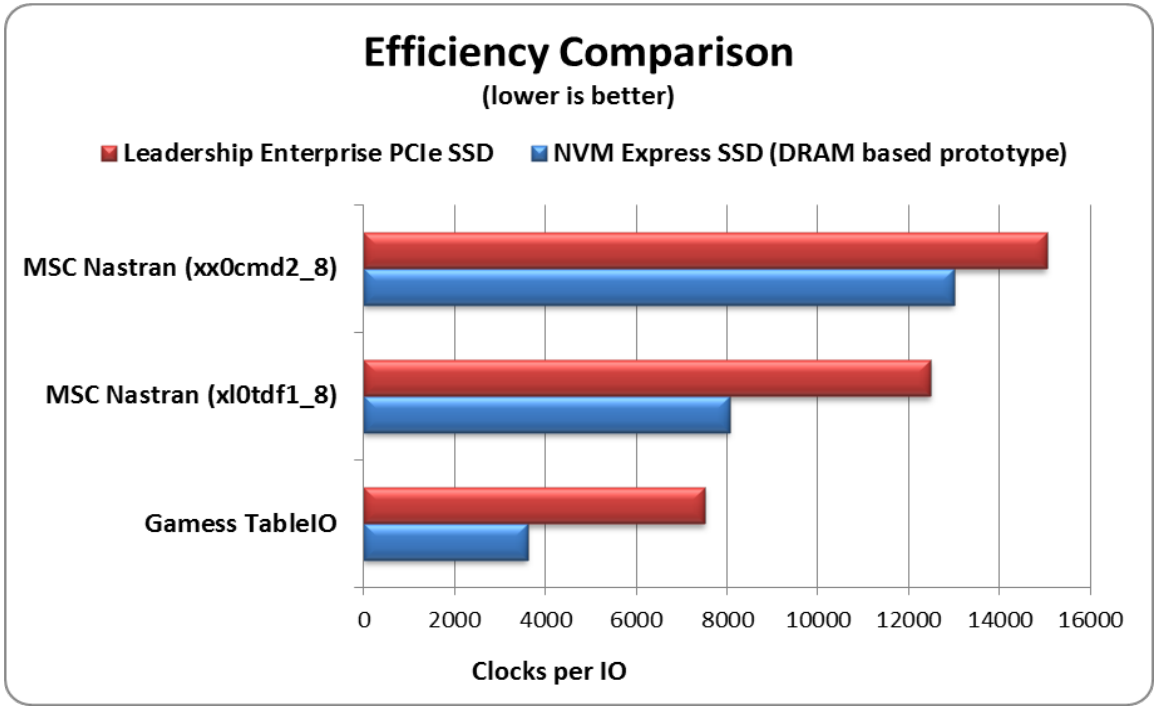
Backup

Performance Efficiency

	AHCI	nvm EXPRESS
Uncacheable Register Reads Each consumes 2000 CPU cycles	4 per command 8000 cycles, ~ 2.5 µs	**0** per command
MSI-X and Interrupt Steering Ensures one core not IOPs bottleneck	No	Yes
Parallelism & Multiple Threads Ensures one core not IOPs bottleneck	Requires synchronization lock to issue command	No locking, doorbell register per Queue
Maximum Queue Depth Ensures one core not IOPs bottleneck	1 Queue 32 Commands per Q	64K Queues 64K Commands per Q
Efficiency for 4KB Commands 4KB critical in Client and Enterprise	Command parameters require two serialized host DRAM fetches	Command parameters in one 64B fetch

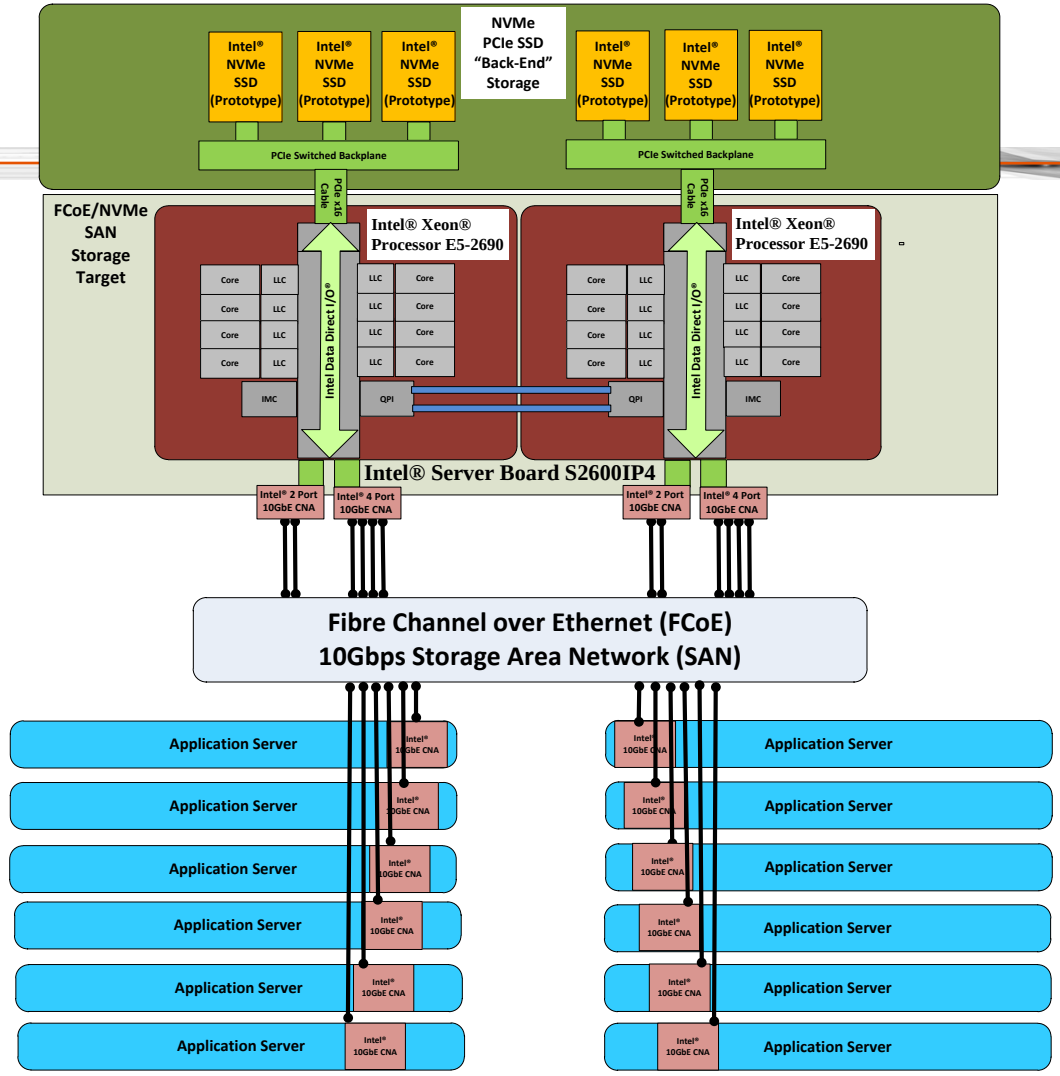
NVM Express Efficiency & Power

- NVM Express prototype delivers lower clocks per I/O while at the same time delivering higher performance on the workloads
- Lower clocks per I/O is a proxy for efficiency and lower power – the CPU & system can go to a sleep state more quickly



Charts compare NVM Express (NVMe) and Leadership Enterprise PCIe *SSD. NVMe utilized DRAM to push protocol to limits. Leadership Enterprise PCIe SSD utilizes NAND making runtime comparisons inappropriate. Gamess TableIO workload is a computational test.

NVM Express (NVMe) in a SAN



- IDF demo combined NVMe with existing ingredients
- The performance of direct attached (DAS) NVMe SSDs married to a Fiber Channel over Ethernet SAN
- Next generation SAN is possible today by use of highly efficient interfaces

- Storage target configuration: Intel® S2600IP4 Server Board, Intel® Xeon® Processor E5-2690 2.9GHz, 8-16GB DDR3 1033 DIMMs, RH EL-6.2 – 3.3.0-RC1 kernel, TCM storage target, , 4 Intel® Ethernet Server Adapter X520 (10 Gbps CNA).
- Initiator configuration: 12 initiators: Intel® Xeon® Processor 5650 2.67GHz, RH EL-6.2 – 3.3.0-RC1 kernel.
- Test configuration: (per initiator) Linux fio V21.0.7, 4K Random Read, QD=8, Workers=16, 8 FCoE LUNs.

SAN with NVMe: 3.1 Million 4K IOPs on 120Gbps FCoE