OPENFABRICS ALLIANCE

nvm EXPRESS™

NVM Express Over Fabrics

11th ANNUAL INTERNATIONAL OPENFABRICS SOFTWARE DEVELOPERS' WORKSHOP

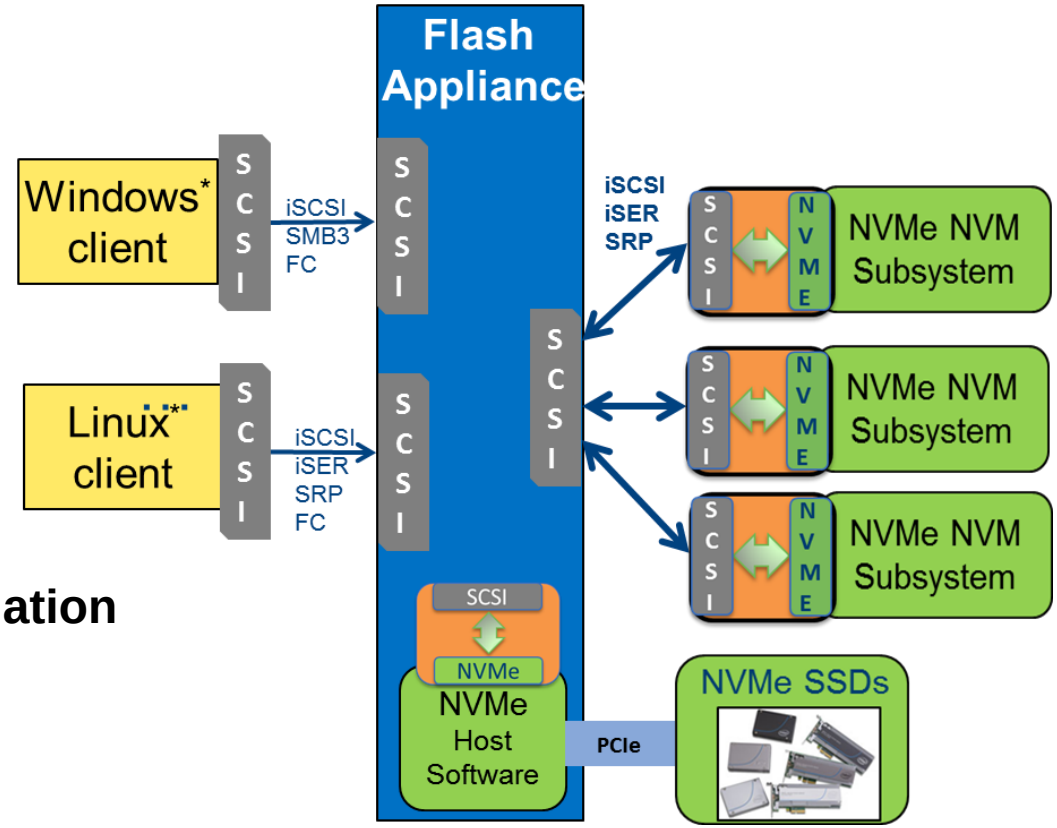
Dave Minturn, Intel Corp.

OFADevWorkshop

NVMe in Fabric Environments

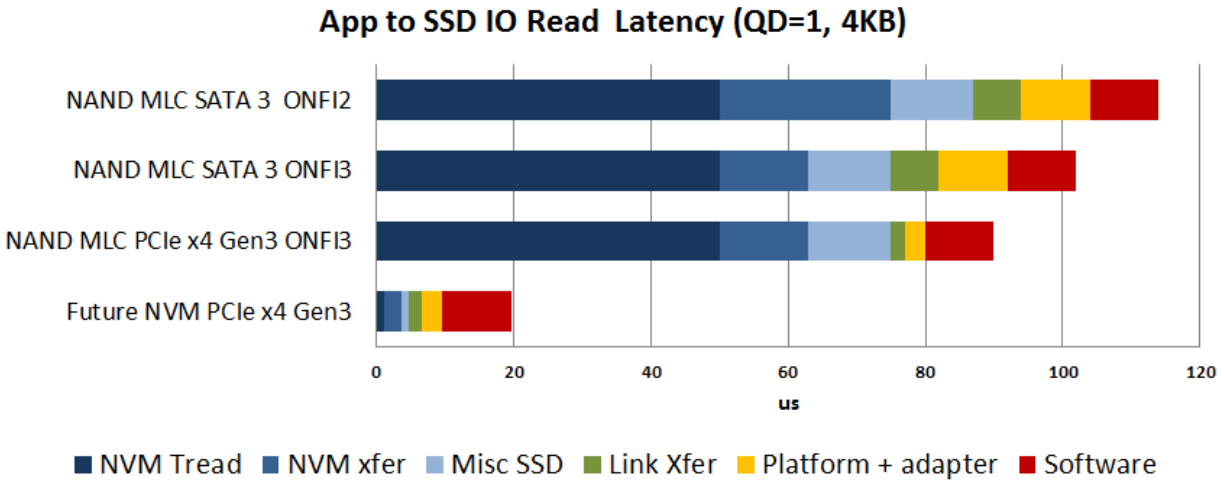
- A primary use case for NVMe PCIe SSDs is in an all flash appliance
- Hundreds or more SSDs may be attached – too many for PCIe based attach scale-out
- Concern: Today remote SSD scale-out over a fabric attach uses SCSI based protocols:

Requiring protocol translation



Desire best performance and latency from NVMe SSD investment over fabrics like Ethernet with RDMA (iWARP,RoCE), InfiniBand™, and Intel® Omni-Path Architecture

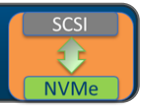
Realizing Benefit of Next Gen NVM over Fabrics



- PCIe NVMe SSD latency may be < 10 µs with Next Generation NVM
- Using a SCSI-based protocol for remote NVMe access ads over 100 µs* in latency
- Usage models require efficient write mirroring of PCIe Next Gen NVMe SSDs over fabrics

*Source: Intel measurements.

Concern: Low latency of Next Gen NVM lost in (SCSI) translation.



Why NVMe over Fabrics?

OPENFABRICS ALLIANCE

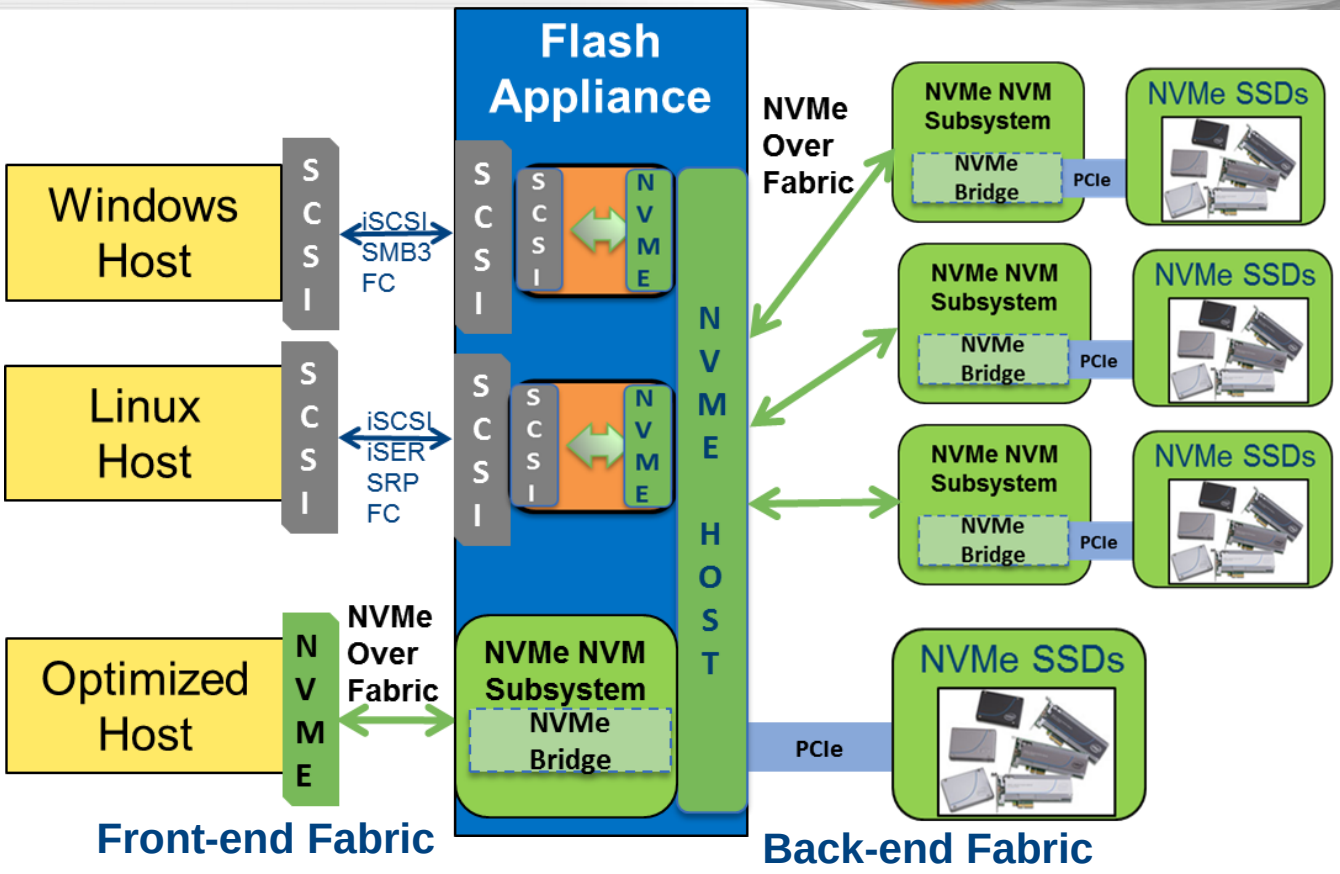
Simplicity, Efficiency and End-to-End NVMe Model

- Simplicity of protocol enables hardware automated I/O Queues – NVMe transport bridge
- No translation to or from another protocol like SCSI (in firmware/software)
- Inherent parallelism of NVMe multiple I/O Queues is exposed to the host
- NVMe commands and structures are transferred end-to-end
- Maintains the NVMe architecture across a range of fabric types
- Maintains architecture and software consistency between fabric types by standardizing a common abstraction and encapsulation definition

Performance Goal:

Make remote NVMe access over fabrics equivalent to local PCIe attached NVMe, within ~ 10 µs latency.

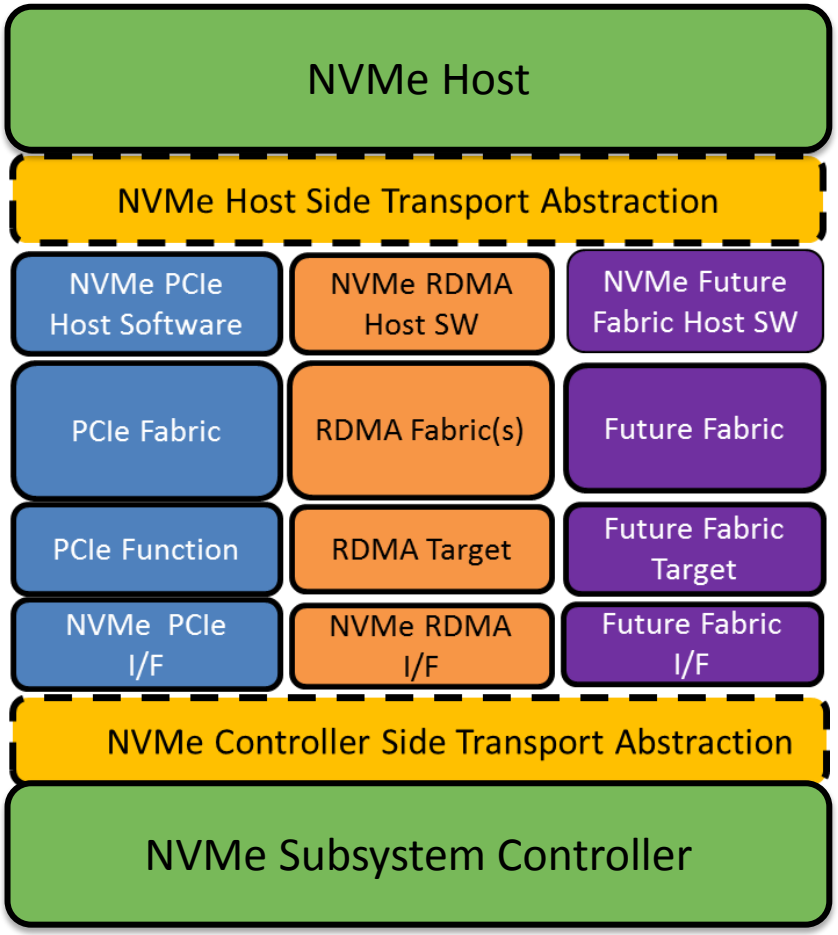
End-to-End NVMe over Fabrics



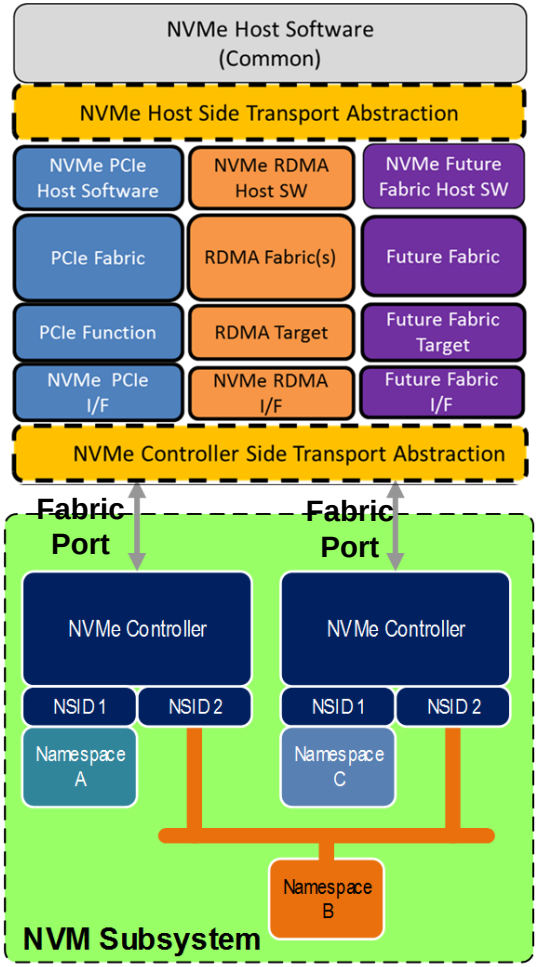
Extend efficiency of NVMe over Front and Back-end Fabrics
Enables efficient NVMe end-to-end model (Host<->NVMe PCIe SSD)

NVMe over Fabrics Architecture
(Standard in definition within nvmexpress.org)

- Maintains a consistent NVMe Host-Controller Architecture Model across all fabric types
  - Common NVMe command set
  - NVMe Multi-queue model
- Transport abstraction layer that enables NVMe over multiple fabric types
  - Fabric agnostic NVMe command and completion encapsulation
  - Fabric-oriented NVMe command data buffer descriptors
- Non-PCIe fabric definitions
  - RDMA family fabrics; Ethernet RDMA (iWARP and RoCE) and InfiniBand™
  - NVMe over Fibre Channel proposed in T11.org



NVMe Architecture is not just for PCIe

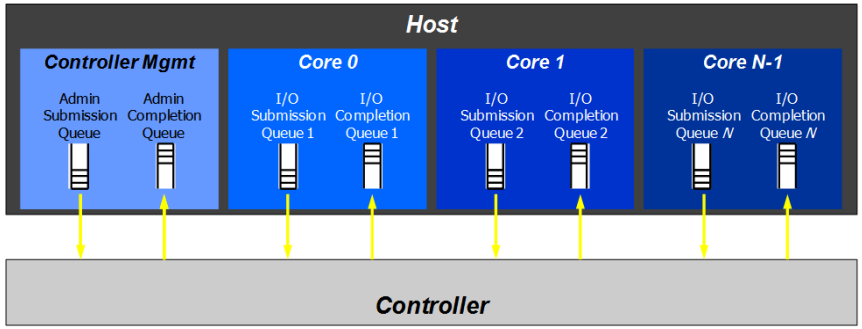


Version 1.2 NVMe Architecture definition is already well suited for non-PCIe Fabric configurations

- **NVM Subsystem Architecture**
  - Multiple NVMe Controllers and fabric ports
  - Multi-path I/O and multi-host support
- **NVMe Namespace Architecture (Pool of logical blocks)**
  - Multiple namespaces per NVM subsystem
  - Can be shared by multiple NVMe Controllers
  - Namespace management
  - Namespace reservations
- **NVMe Controller multi-queue host interface**
  - Administrative and multi-IO queue model
  - Simple command set, optimized for NVM
  - SGL based host and controller buffer descriptors

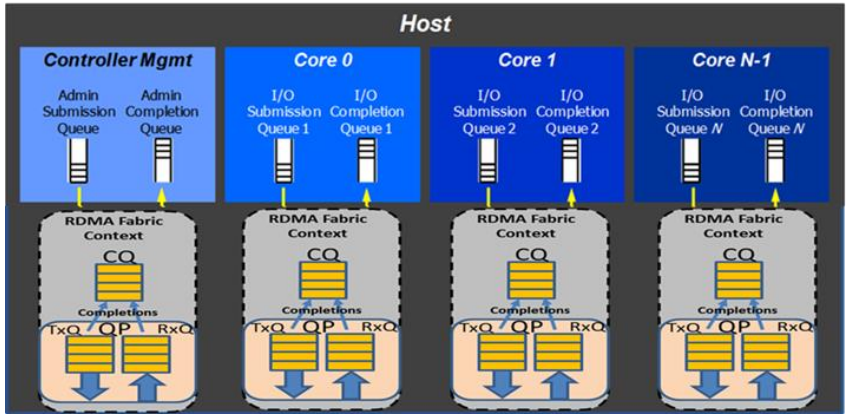
NVMe over Fabrics extends the NVMe Architecture elements over multiple fabric types

NVMe Multi-Queue Host Interface



- NVMe Submission and Completion Queues are aligned to CPU cores
- No inter-CPU software locks
- Per CQ MSI-X interrupts enable source core interrupt steering

NVMe multi-queue interface maps nicely onto multi-queue enabled high-performance fabrics such as the RDMA queue pair model

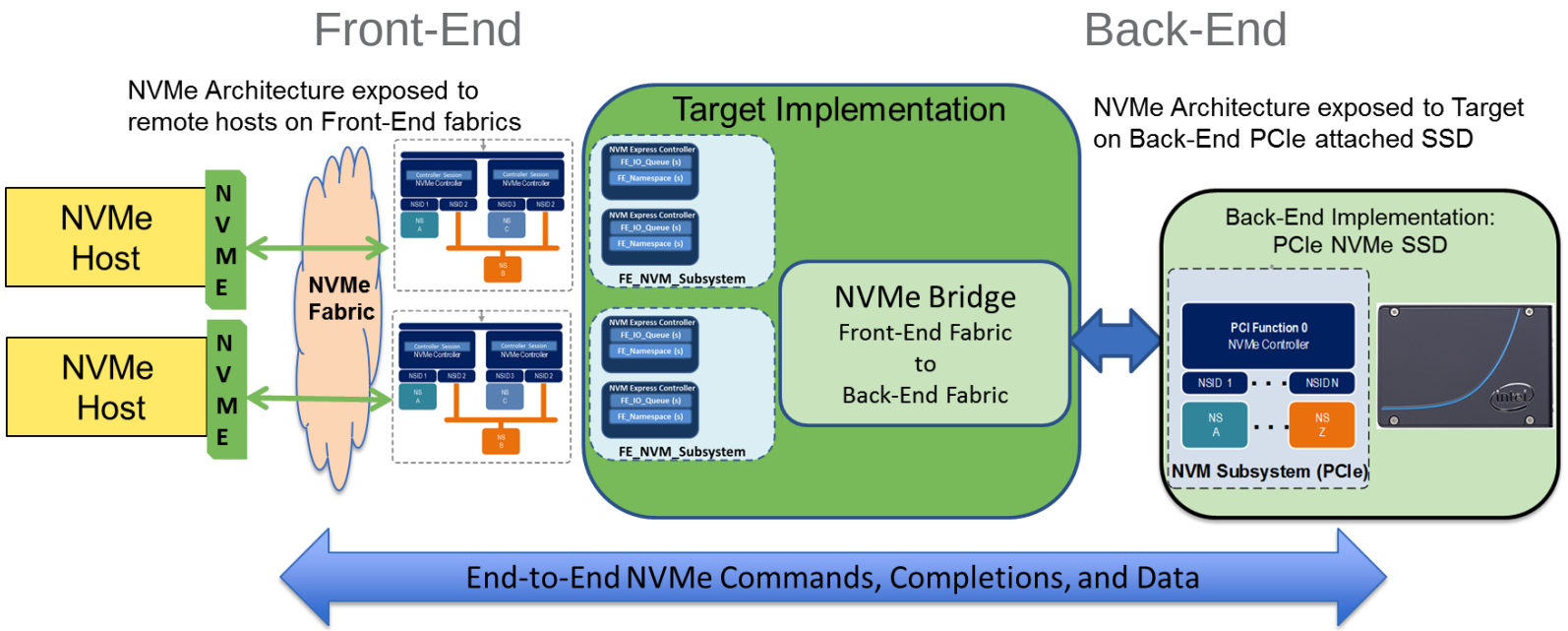


(*Possible RDMA QP mapping shown)

- Retains NVMe SQ/CQ CPU alignment
- No inter-CPU software locks
- Source core interrupt steering retained by using RDMA Event Queue MSI-X interrupts

NVMe over Fabrics retains the NVMe host interface model for host software efficiency and consistency

NVMe over Fabric Target



- Exposes NVMe Subsystems to hosts on Front-End NVMe fabric
- Bridges to Back-End NVMe Subsystems (PCIe SSDs)
- NVMe over Fabric Targets are implementation specific
  - Server software based target implementations
  - Single/multiple drive device based target implementations

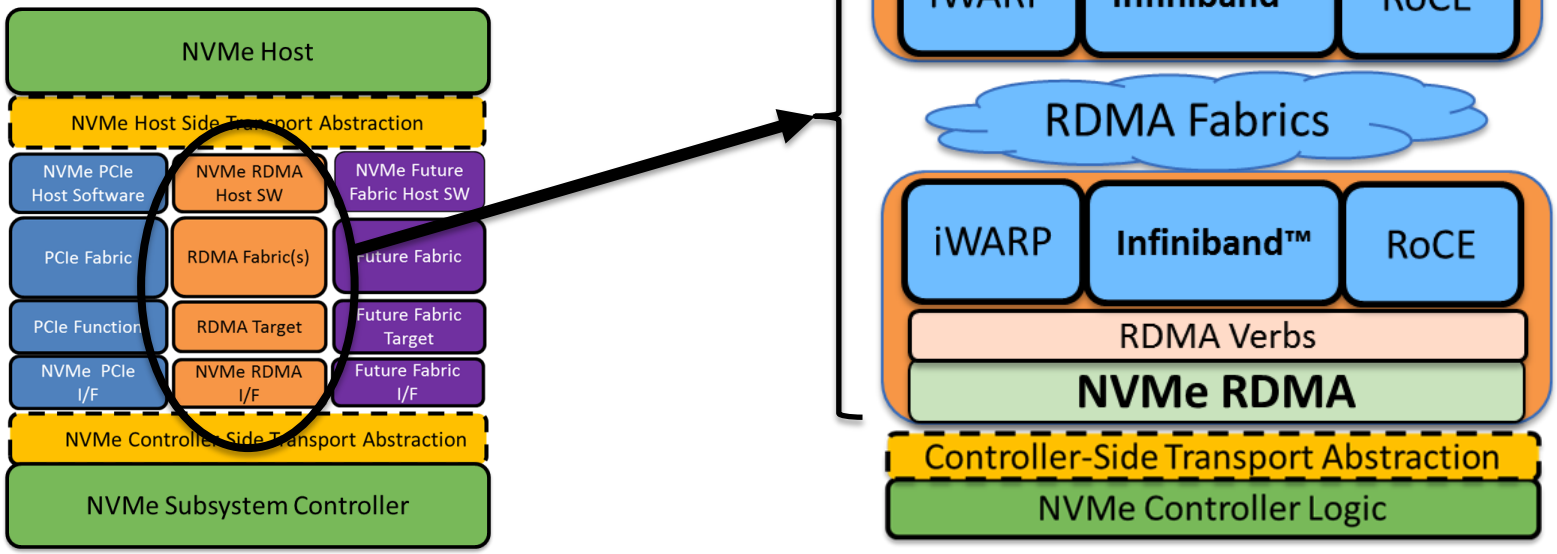
NVMe over RDMA Fabrics

(*Standard in definition within nvmexpress.org)

OPENFABRICS ALLIANCE

NVMe over RDMA Fabric

- Upper Level RDMA Block Storage Protocol
- Layered over a common set of RDMA Verbs
- Imperative to support all RDMA provider types
  - Infiniband™
  - Ethernet (iWARP and RoCE)



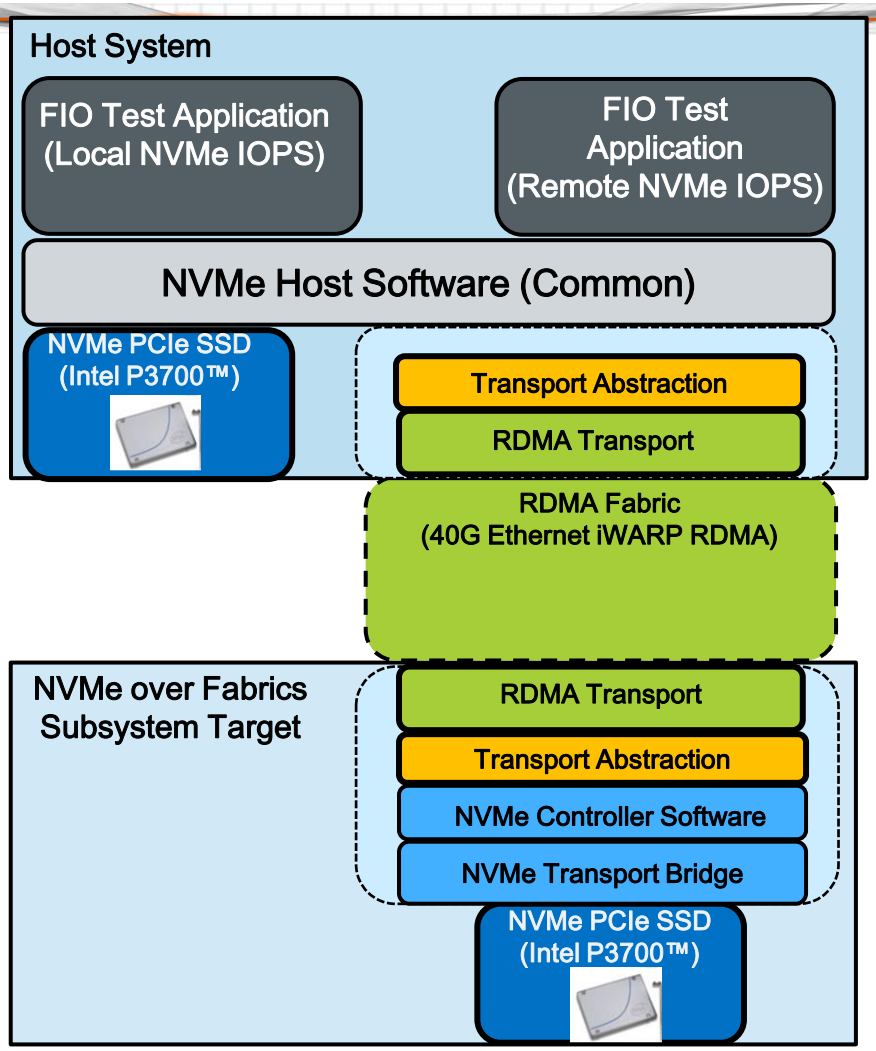
NVMe over Fabrics concept shown at Intel Development Forum 2014

OPENFABRICS ALLIANCE

NVM Express over 40GbE iWARP RDMA

- Recall: Goal is remote NVM Express (NVMe) equivalent to local NVMe, up to 10 µs added latency
- Prototype delivers 450K IOPs for ***both*** the local and remote NVMe PCIe devices
- Remote NVMe adds ***8 µs*** latency versus local NVMe access
- Demonstrates the efficiency of NVMe End2End NVMe Target software running on one CPU core (two SMT threads) at 20% Utilization

NVMe over Fabrics Standard in Definition
Get involved through NVM Express Org.



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.

Questions?

Thank You

OPENFABRICS ALLIANCE

ANNUAL INTERNATIONAL OPENFABRICS SOFTWARE DEVELOPERS' WORKSHOP