OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

NVME™ OVER FABRICS

Presented by Phil Cayton
Intel® Corporation

April 6th, 2016

NVME OVER FABRICS

- **NVM Express™* Organization**
- **Scaling NVMe in the datacenter**
- **Architecture / Implementation Overview**
- **Standardization and Enabling**

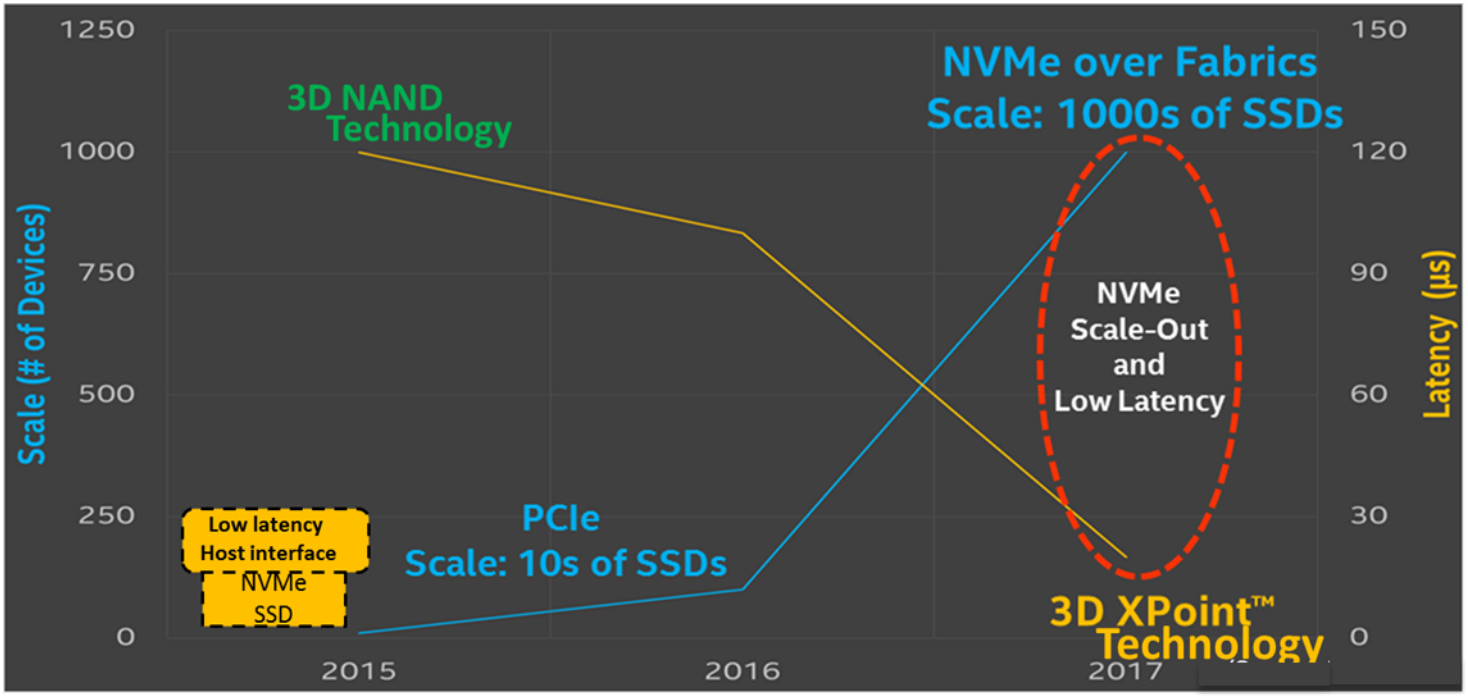
NVME OVER FABRICS

NVM Express Organization

- 80+ companies strong and growing
- Workgroups:
  - Technical – Specifications
  - Driver – Linux™* Host and Target fabrics driver
  - Marketing – NVMe awareness
- Learn more at nvmexpress.org

NVME OVER FABRICS

Evolution of Non-Volatile Storage in the Datacenter

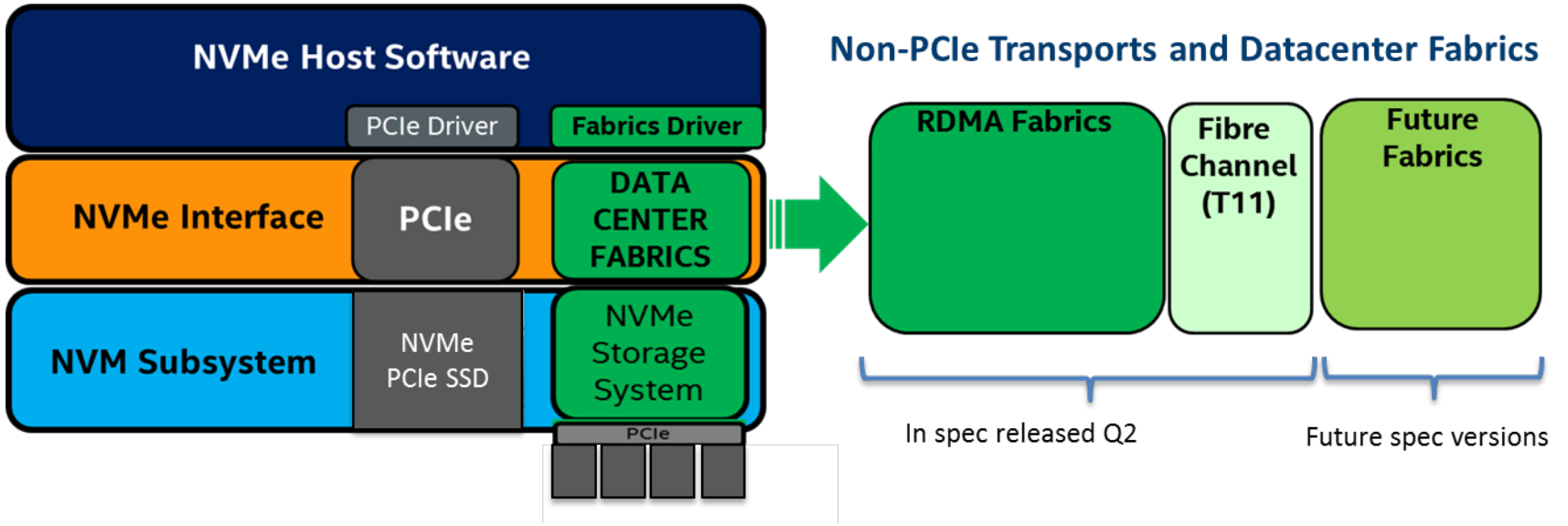


- **Effective scaling of PCIe* attached NVMe* is limited to 255/system**
- **Datacenter requires scaling out to 100's or 1000's of NVMe SSDs**
- **Traditional scaleout adds translations between Hosts and NVMe SSD**
- **NVMe over Fabrics extends NVMe and enables SSD Scale-Out & Low-Latency I/O needed by the datacenter at near local speeds**

NVME OVER FABRICS

Industry standard definition of NVMe over Datacenter Network Fabrics

- **Shares same base architecture and NVMe Host Software as PCIe**
- **Specification defines interface supporting multiple network fabrics**

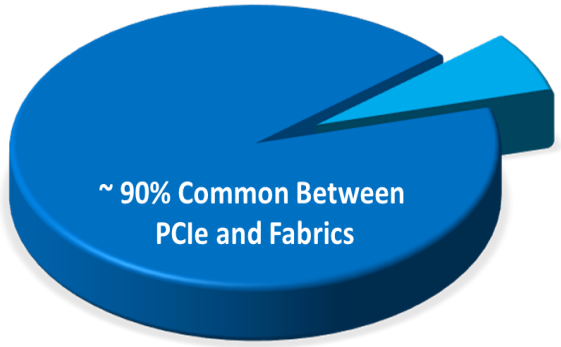


NVMe over Fabrics specification defines how NVMe is extended to support new network fabrics

NVME OVER FABRICS

Commonality Between NVMe on PCIe and on Fabrics

- **The vast majority of NVMe architecture is leveraged as-is for Fabrics**
  - NVMe Multi-Queue Host Interface, Subsystem, Controllers, Namespaces, and Commands
  - Allows for use of common NVMe Host software with very thin fabric dependent layers
- **Primary differences reside in the discovery and queuing mechanisms**



Differences	PCI Express® (PCIe)	Fabrics
Identifier	Bus/Device/ Function	NVMe Qualified Name (NQN)
Discovery	Bus Enumeration	Discovery and Connect commands
Queuing	Memory-based	Message-based
Data Transfers	PRPs or SGLs	SGLs only, added Key

NVME OVER FABRICS

Message Based Queueing (via encapsulation)

- **Fabric Capsules are messages with "encapsulated" NVMe content**

Host to NVMe Controller

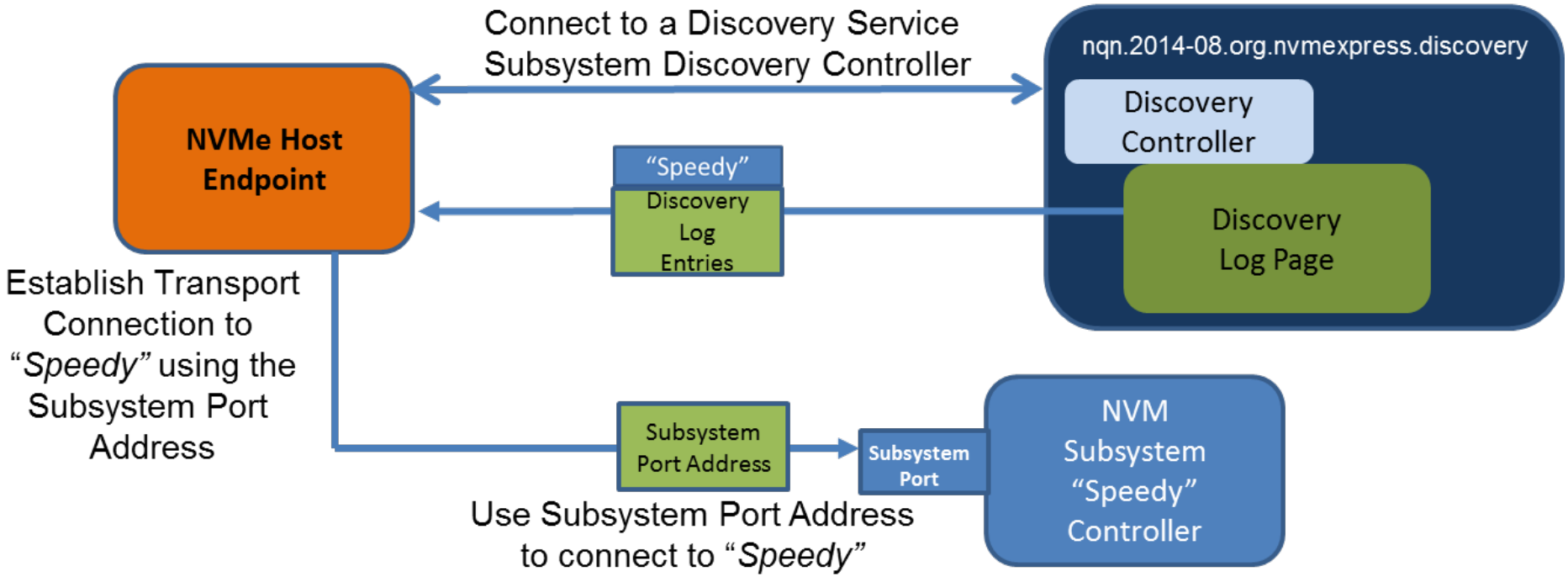
NVMe Controller to Host

- **Data Section may be sent <u>either</u> within the Capsule <u>or</u> via a fabric type dependent data transfer mechanism (RDMA_READ/ RDMA_WRITE)**

Data transfer to/from host resident buffer

NVME OVER FABRICS

Fabric Subsystem Discovery



Discovery Log Entries

- Information needed to establish connections to NVM Subsystems
- Accommodates multiple transports and address types

NVME OVER FABRICS

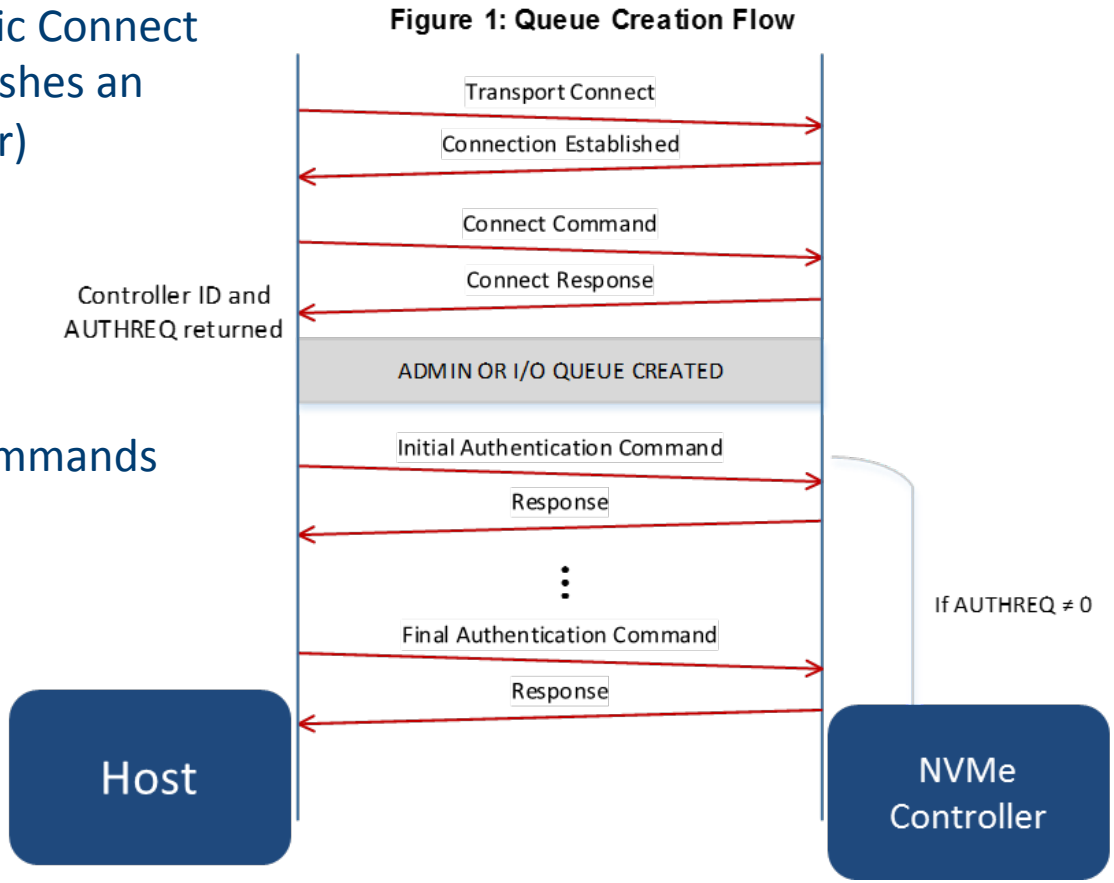
Queue Creation using Fabric Connect Command

1) Create a fabric-dependent transport connection

2) Send a Command Capsule with Fabric Connect Operation (AdminQ Connect establishes an "association" to an NVMe Controller)

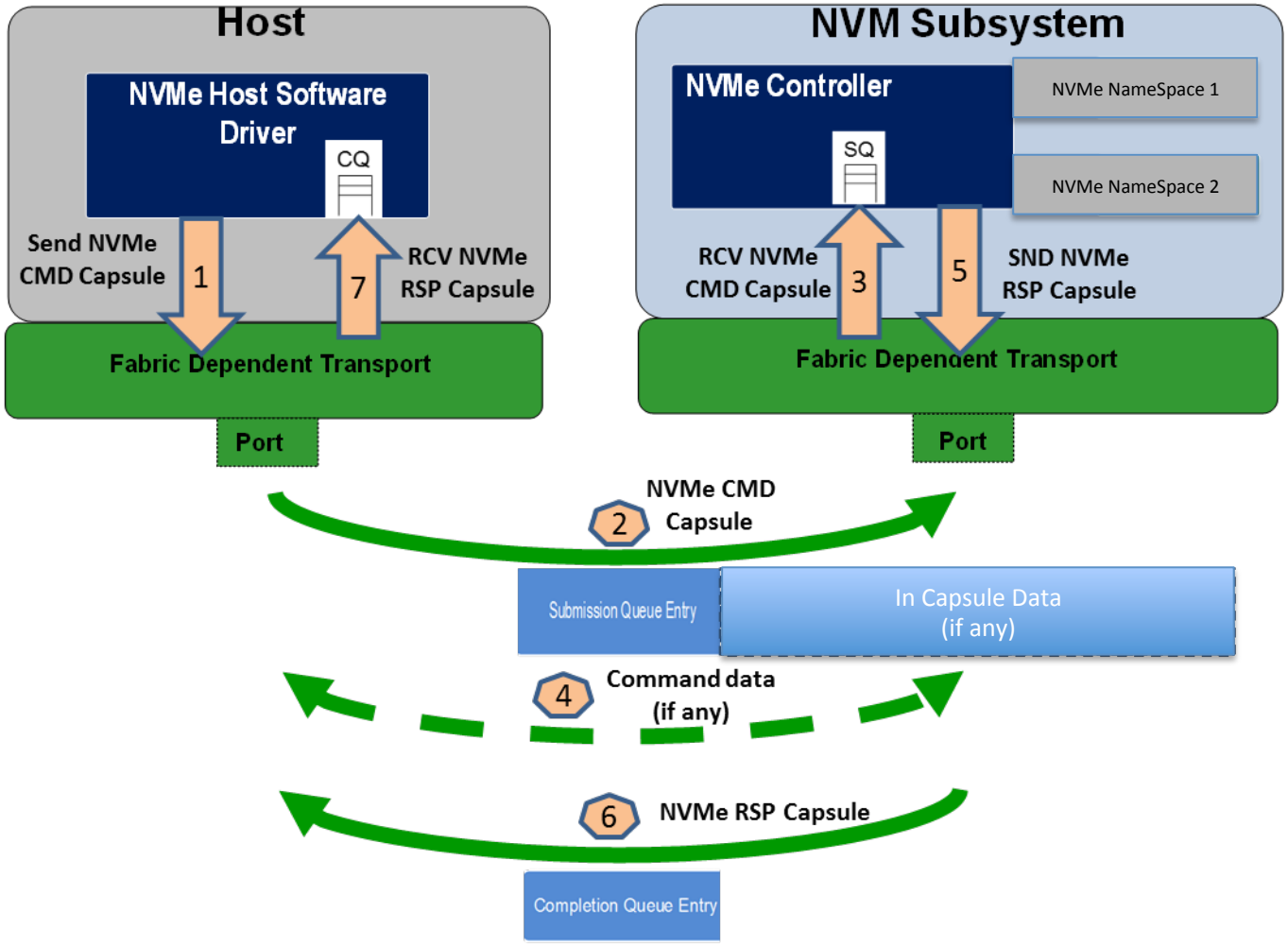
3) Send <any> Authentication Fabric Commands

4) AdminQ or IOQ Ready for NVMe Commands



NVME OVER FABRICS

Capsule Exchange Example

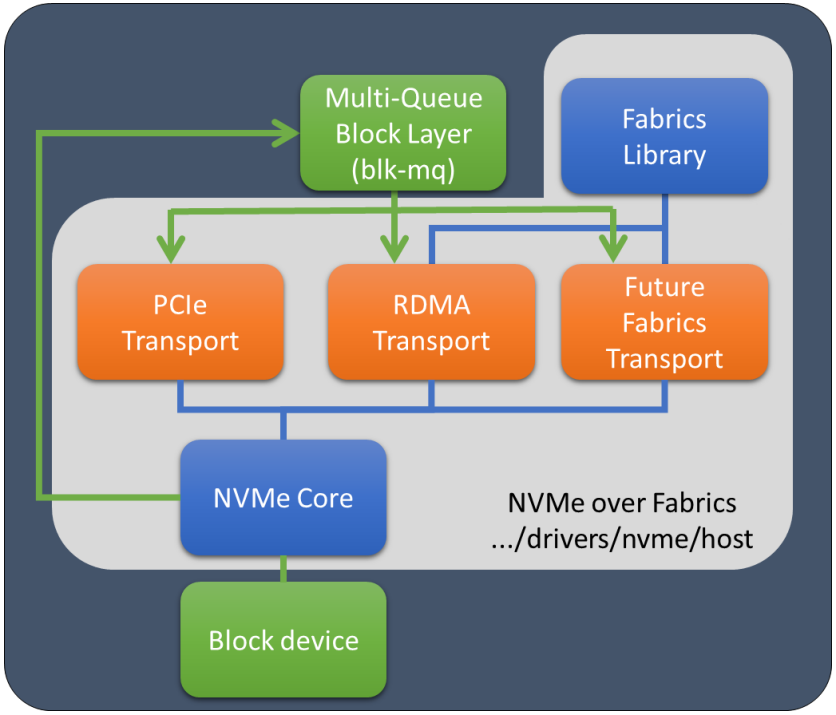


NVME OVER FABRICS

Host Components

- **Architecture**
  - Separation of NVMe Core Logic and Transport
  - Common NVMe Core across all transports: OS and Block interface
  - Common NVMe Fabric agnostic functions: Capsules, Properties, Connect, and Discovery
  - Thin transports with limited NVMe awareness that can interoperate with multiple targets
- **Implementation**
  - Separate Driver Components
  - Components interface with Multi-Q Block layer
  - Pluggable Transports

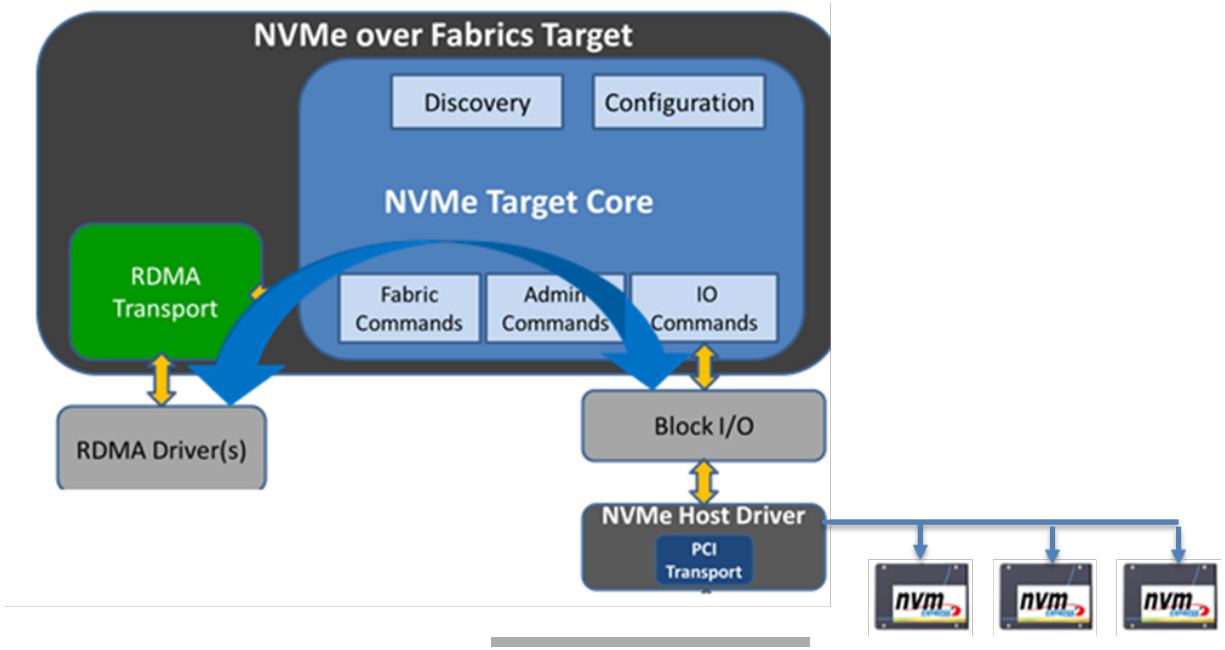
Retains efficiency of PCIe NVMe access over fabrics



NVME OVER FABRICS

Target Components

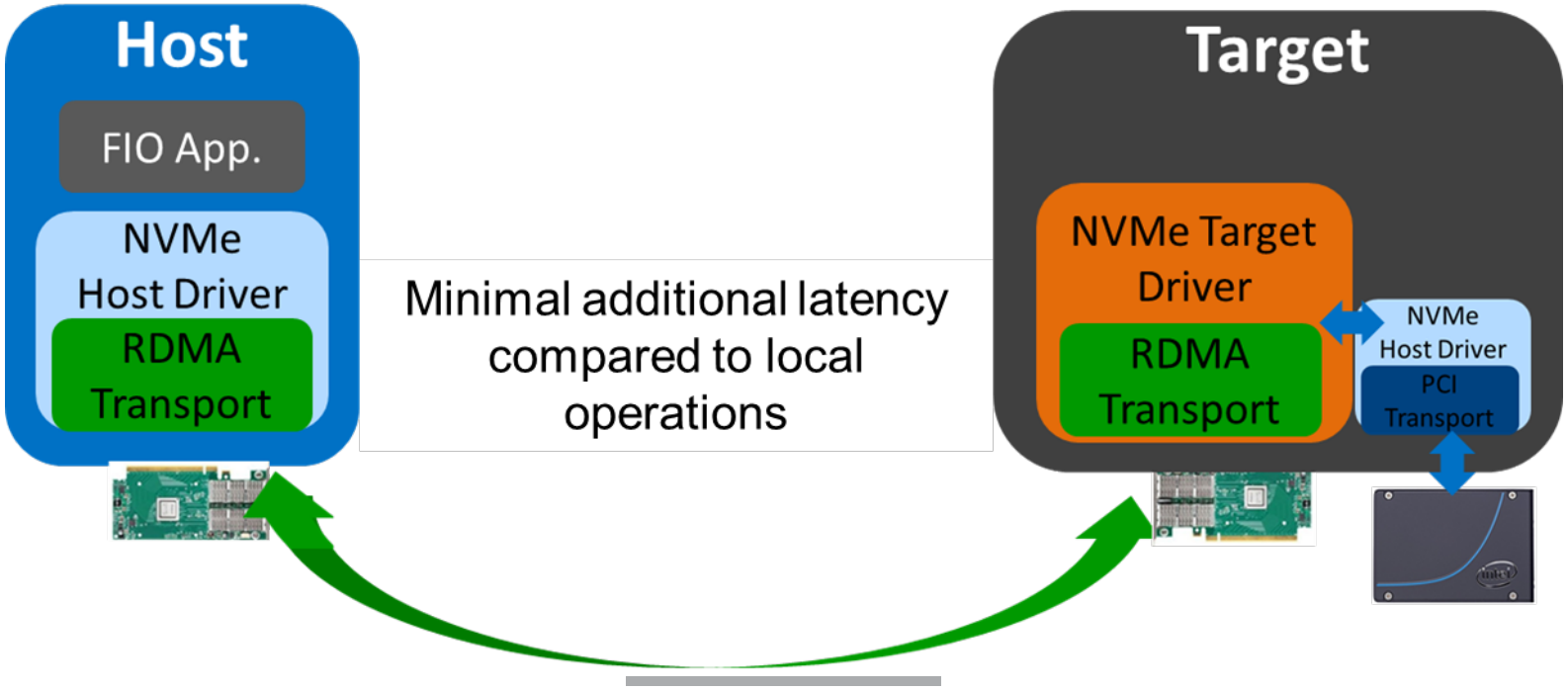
- **Architecture**
  - Virtual NVMe subsystem representation of PCIe SSD Subsystem
  - Separate NVMe subsystem core components and transports
  - Discovery Subsystem and controllers
- **Implementation**
  - Creates logical NVMe Subsystems and Controllers that are presented to Hosts
  - NVMe Namespaces are logically mapped to physical block devices



NVME OVER FABRICS

Host and Target Stack Status

- **Linux Host and Target Kernel drivers operational**
  - Target configured with NVMe PCIe SSD
  - To ensure fabrics transport is fabric agnostic it has been tested on InfiniBand™*, iWARP, RoCE RDMA adapters
  - Host has been tested with multiple target implementations
- **Drivers are still being performance tuned**



NVME OVER FABRICS

NVMe over Fabrics Standardization and Enabling

- **NVMe over Fabrics Specification complete**
  - In release candidate final review
  - Available publicly on nvmexpress.org: 05/16

- **Host and Target drivers will available in upstream Linux kernel**
  - Upstream Kernel under drivers/nvme/{host/target}
  - Available: shortly after specification released in ~4.8
  - Fabrics supported:
    - Local: PCIe
    - RDMA: Fabric agnostic kernel verbs transport
    - Fibre Channel (under development)
    - Other fabrics transports under evaluation / pathfinding (e.g., kFabric)

When the specification and Host and Target code is available
Please download, test with your fabric hardware and environments

LEGAL DISCLAIMER & OPTIMIZATION NOTICE

- **No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade. This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps. The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.**
- **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.**
-
- ***Other names and brands may be claimed as the property of others**

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

OPENFABRICS
ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

Intel®

NVME OVER FABRICS

Target Components