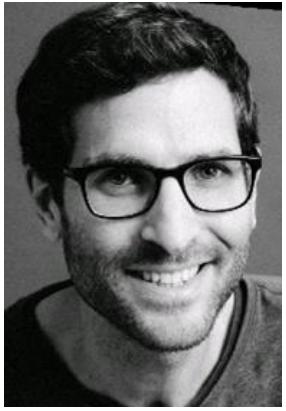
An Overview of the Refactored NVMe Transport Specifications – PCIe®, RDMA, and TCP

Sponsored by NVM Express organization, the owner of NVMe® specifications

Speakers

John Geldman
Director, SSD Industry Standards
KIOXIA

Curtis Ballard
Distinguished Technologist
Hewlett Packard Enterprise



Sagi Grimberg
CTO
lightbits

Flash Memory Summit

nvm EXPRESS®

Agenda

- Refactoring in NVMe® 2.0 Specifications
- PCIe Transport Specification
- RDMA Transport Specification
- TCP Transport Specification

Refactoring the transport specifications

How we started: A short NVM Express® story

- NVM Express 1.0 was released in 2011
  - The NVM Express specification defined a register level interface for host software to communicate with a non-volatile memory subsystem over PCI Express®
- NVM Express 1.2.1 and NVMe-oF™ 1.0 were released in 2016
  - NVMe 1.2.1 added NVMe over Fabrics as a new type of NVMe transport, with its own specification
  - Examples included: Ethernet, InfiniBand™, and Fibre Channel
  - Support requirements for features and functionality differed for PCI Express and Fabrics transports

Transport Spec Refactoring: The What and the Why

- In 2021, NVMe® 2.0 “Refactored” the NVM Express® Specifications
- Refactoring the transport specifications
  - Cleanly separate common NVM Express functionality from PCIe® specific
    - The NVM Express specification was initially PCIe technology only
    - PCIe specific terminology, references, etc. were spread throughout
  - Cleanly merge NVMe-oF common functionality into the base specification
  - Cleanly separate transport specific details into their own specifications
- Result: Independent transport specifications able to remain stable or be updated without changing unrelated specifications

NVM Express® PCIe® Transport Specification

PCI Express®, the Original NVMe® Transport

In 2011, NVM Express was released with PCI Express as its backbone

In 2022, PCIe® architecture is still a popular NVMe transport and has been amazingly efficient (compared to previous storage protocols)

- NVMe technology was developed with an eye for efficient, latency minimized interactions
- NVMe technology's multiple command/submission queue, doorbell, and MSI-X interrupt structures were built with efficiency in mind – this efficiency is still a focus

NVM Express® technology is now relatively independent of PCI Express® versions

NVM Express has been working on removing PCI Express revision specific requirements (those are the domain of PCI-SIG®)

NVM Express has benefited from being able to ride the waves of performance and feature improvements from PCI-SIG

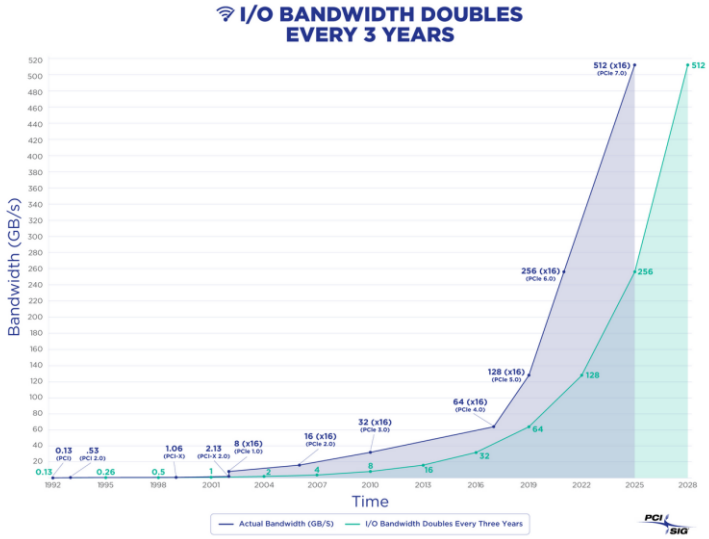
Requirements on the PCI Express interface for NVM Express architecture are focused in two specifications:

- The NVMe® over PCIe specification, initially developed in the refactoring, defines NVMe requirements and behaviors that are specific to the PCIe transport
- The NVMe Management Interface specification defines how the PCIe command set is used for an NVM Express Management Interface

NVMe® technology rides the PCI Express performance ramp

NVMe technology intersected PCI Express technology at Rev. 3.0 PCI Express has continued to evolve:

- Revision 3.0 @ up to 8 GT/s per lane in 2010
- Revision 4.0 @ up to 16 GT/s per lane in 2017
- Revision 5.0 @ up to 32 GT/s per lane in 2019
- Revision 6.0 @ up to 64 GT/s per lane in 2021
- PCI Express has announced the development of Revision 7.0 @ up to 128 GT/s per lane (targeting 2025)



NVMe® technology rides the feature improvements of PCI-SIG®

PCIe 5.0 specification and beyond

- CMA: Content Measurement & Authentication
- IDE: Integrity and Data Encryption
- Combined Power: A reformed power budgeting mechanism (still co-exists with NVMe Power States)

PCIe 6.0 specification and beyond

- FLITs: A new packet format (FLow control unIT), a 256 byte structure with Forward Error Correction
- L0p: a mechanism to bring up and down lane counts without stalling

NVMe® technology rides the feature improvements of PCI-SIG®

PCIe® Technology Ongoing Initiatives (Join PCI-SIG to contribute or for more)

- Trusted Execution Environment (TEE): An environment to set up confidential workloads, isolated from hosting environment
- I3C Basic: (still in proposal) Developing a uniform implementation for PCIe ecosystem (voltages and relationship to SMBus)

NVM Express® RDMA Transport Specification

NVMe® RDMA Spec – Revealing the Simplicity

- NVM Express® RDMA Transport Specification 1.0a
  - 16 pages of simplicity!
- RDMA operations align closely with PCIe® architecture operations
- Pulling RDMA transport requirements into their own specification reveals how directly the NVMe technology model maps to RDMA

RDMA Specification Updates

- The RDMA Transport Specification 1.0a is functionally equivalent to the RDMA transport requirements in the NVMe® over Fabrics 1.1a specification
- Changes include:
  - Incorporates definitions from Infiniband specification for generic RDMA terms
  - Aligns terms used by NVMe technology with RDMA defined terminology
  - Documentation structure and introductory text clauses to support separating into a standalone document

NVM Express® TCP Transport Specification

NVMe®/TCP Spec – Queueing, messaging and specific features

- NVM Express® TCP Transport Specification 1.0b - 35 pages long
  - Setup and Initialization
  - Queueing model
  - Data Transfer
  - Wire format
  - Error handling
  - Data integrity and Security
- Future extensions will introduce new PDU formats

TCP Specification Updates

- Ratified:
  - TLS 1.3 update
  - Inband Authentication (with TLS)
- Related WIP Technical proposals:
  - Authentication Verification Entity for DH-HMAC-CHAP (TP 8019)
  - X.509 certificates for NVMe-oF endpoints and use with TLS (TP 8023)
- Related
  - Automated Discovery of IP Discovery Controllers (TP 8009)
  - NVMe-oF Centralized Discovery Controller (TP 8010)
  - NVMe Boot specification

TCP ecosystem update

- **Linux support is maturing**
  - Various bug reports, new testers, new devices stepping forward
  - userspace toolchains expanding and natively support NVMe/TCP
  - New addition for Inband-auth (pending inclusion)
  - NVMe/TLS (Working PoC from Hannes) - joint work with NFS folks
- **Now supported in VMware!**
  - Native support introduced in 7.0U3
  - Since then a few patch releases, as well as (7.0U3e)
  - Two zero-day partners cooperated with VMware on NVMe/TCP
  - Already running in production!
- *External drivers for windows in the wild, maybe soon inbox?*

Questions?


nvm
EXPRESS®
Architected for Performance