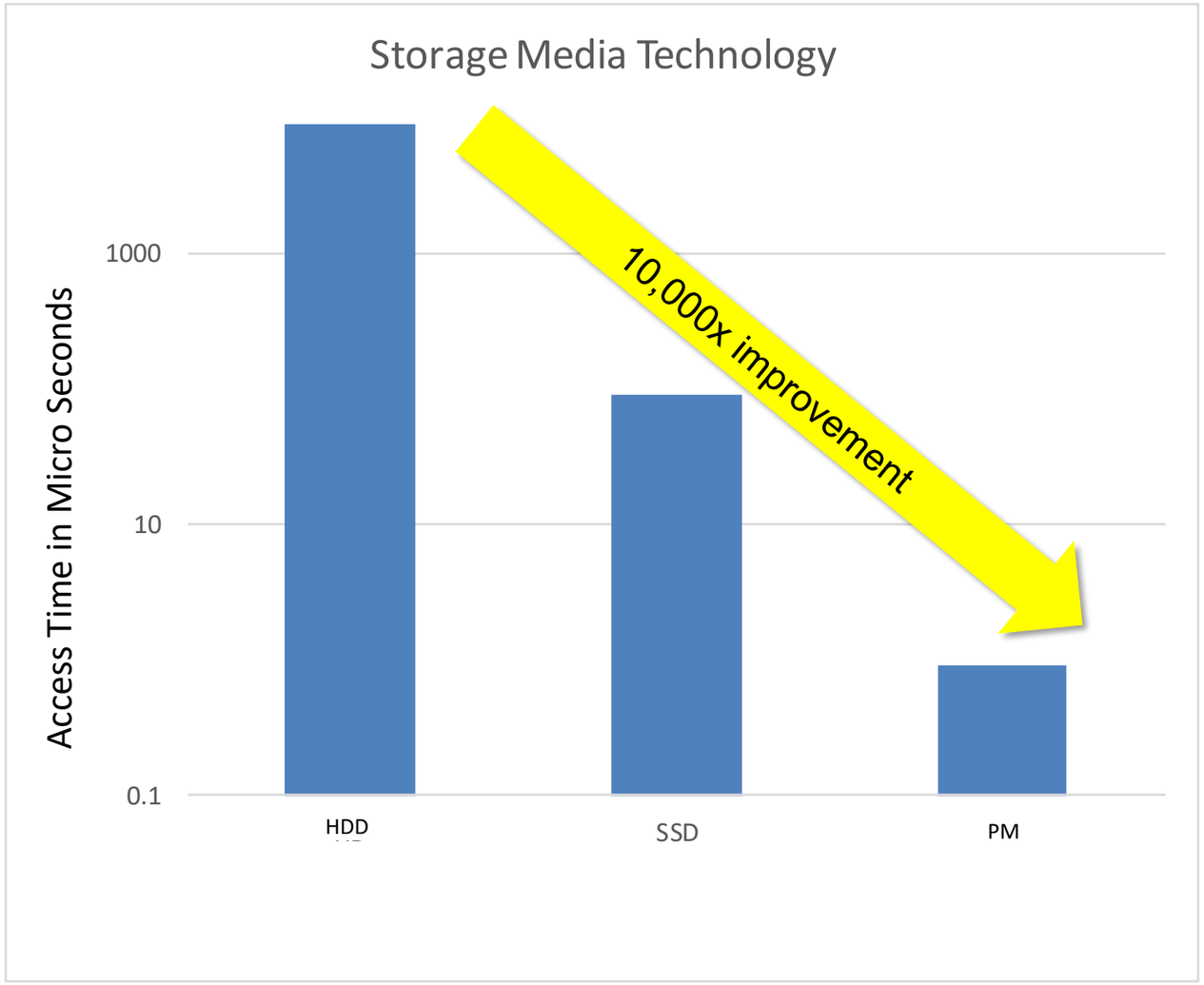
SDC 19
SNIA INDIA
May 23-24, 2019
Bangalore, India
STORAGE DEVELOPER CONFERENCE

NVMe over Fabrics Demystified

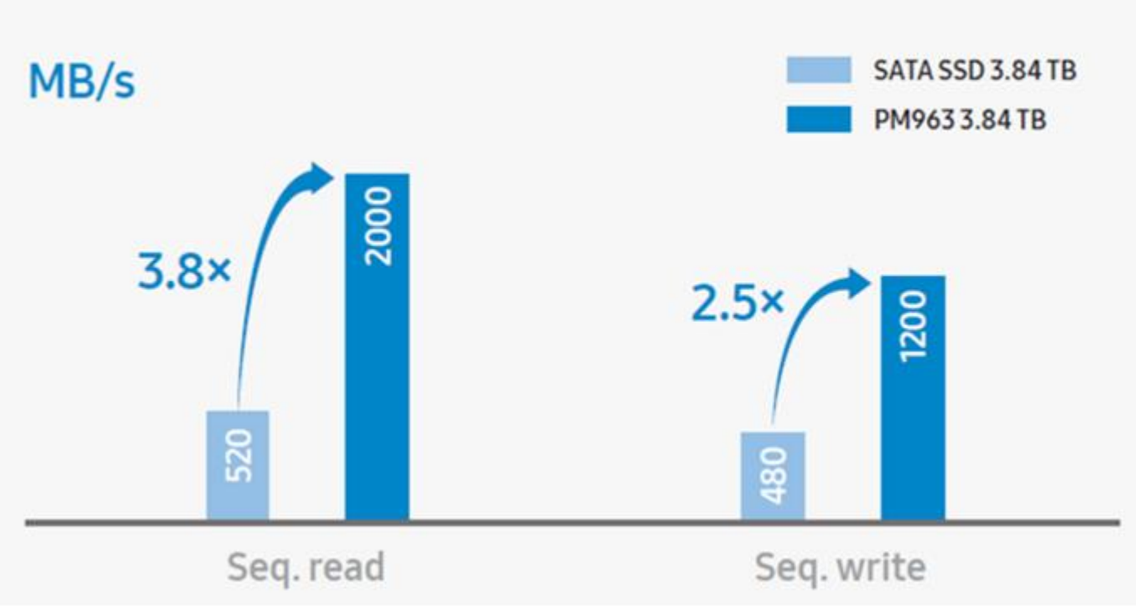
Rob Davis
Mellanox

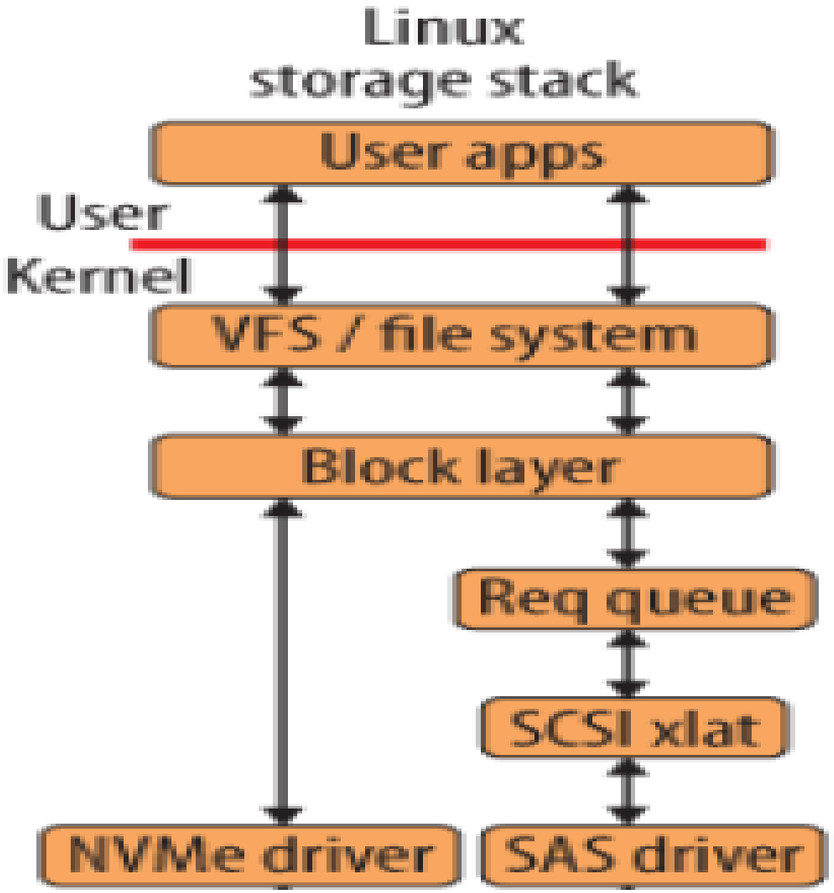
Why NVMe over Fabrics?



NVMe Technology

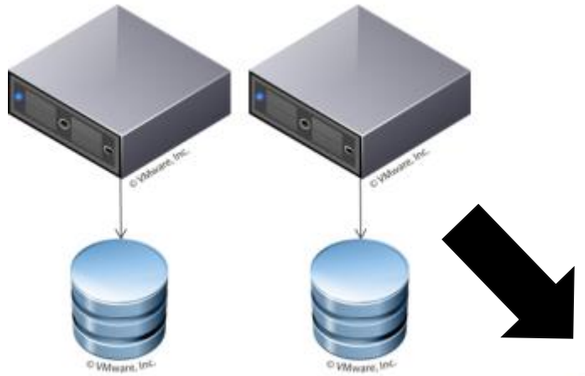
- Optimized for flash and PM
  - Traditional SCSI interfaces designed for spinning disk
  - NVMe bypasses unneeded layers
- NVMe Flash Outperforms SAS/SATA Flash
  - +2.5x more bandwidth, +50% lower latency, +3x more IOPS

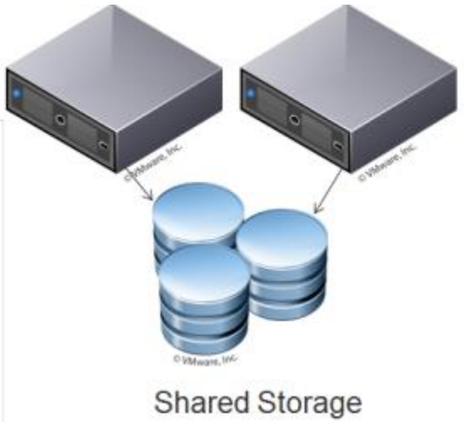


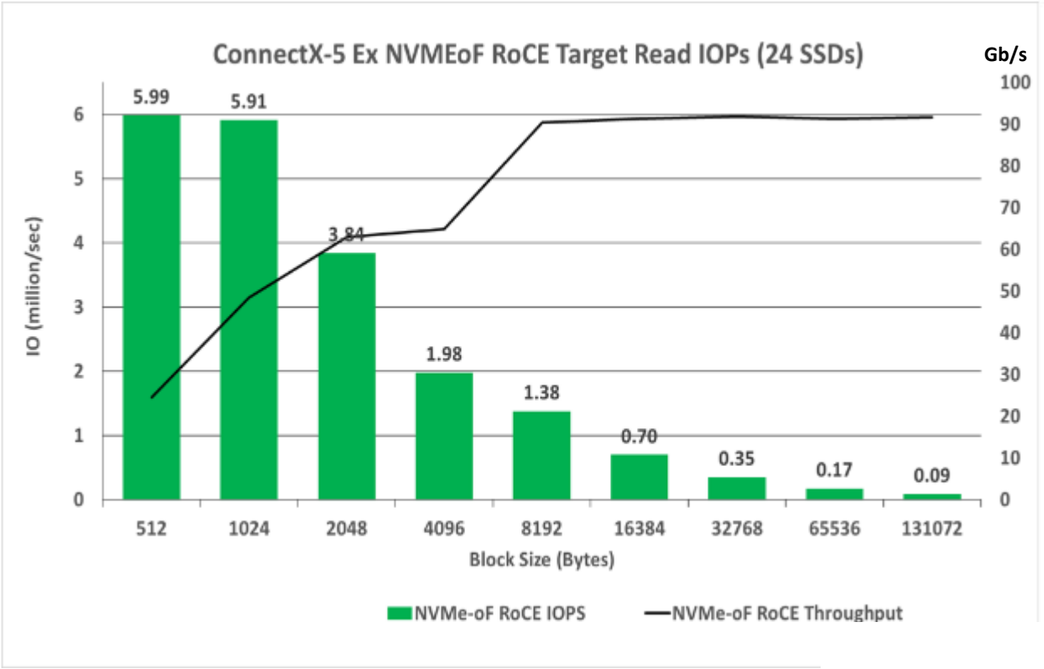


“NVMe over Fabrics” was the Logical and Historical next step

- Sharing NVMe based storage across multiple servers/CPUs was the next step
  - Better utilization: capacity, rack space, power
  - Scalability, management, fault isolation
- NVMe over Fabrics standard
  - 50+ contributors
  - Version 1.0 released in June 2016
- Pre-standard demos in 2014
- Able to almost match local NVMe performance

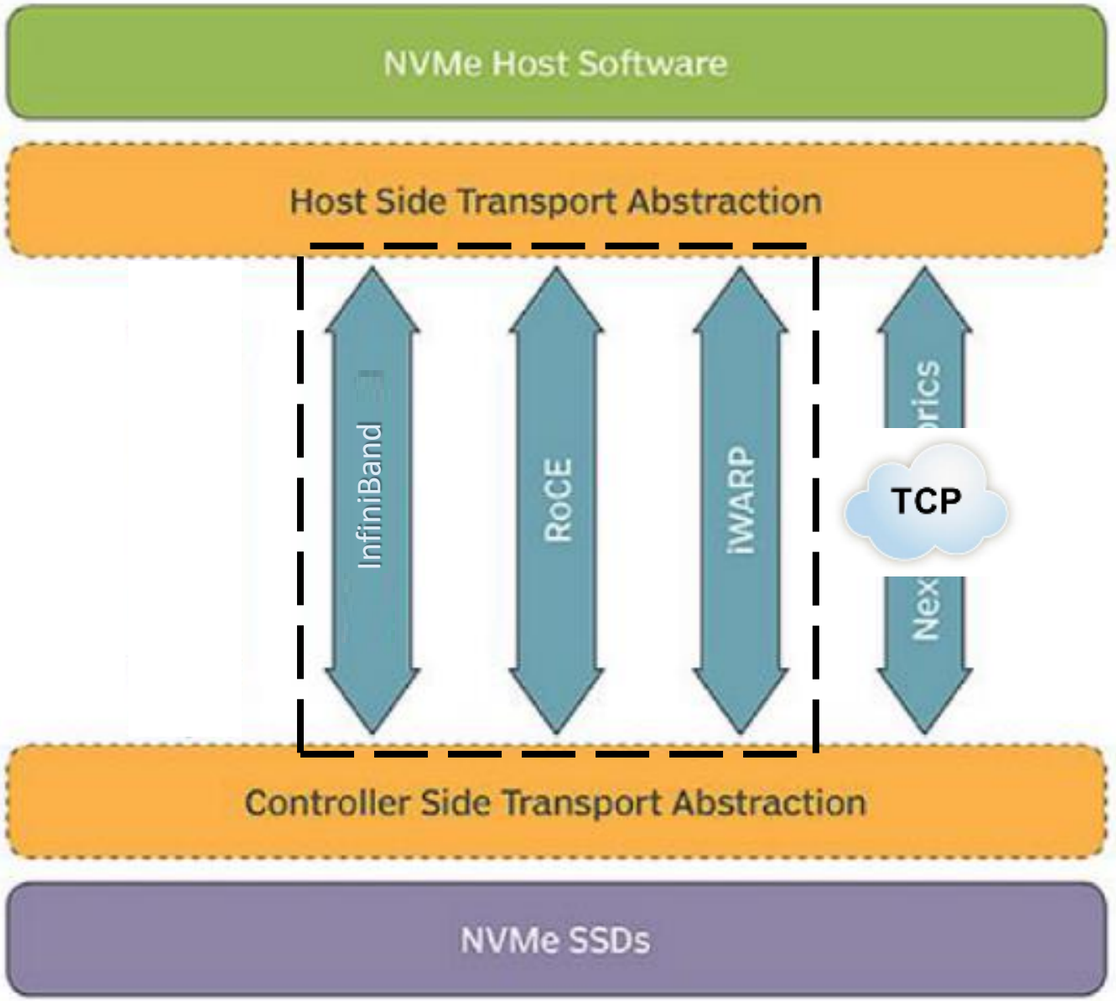






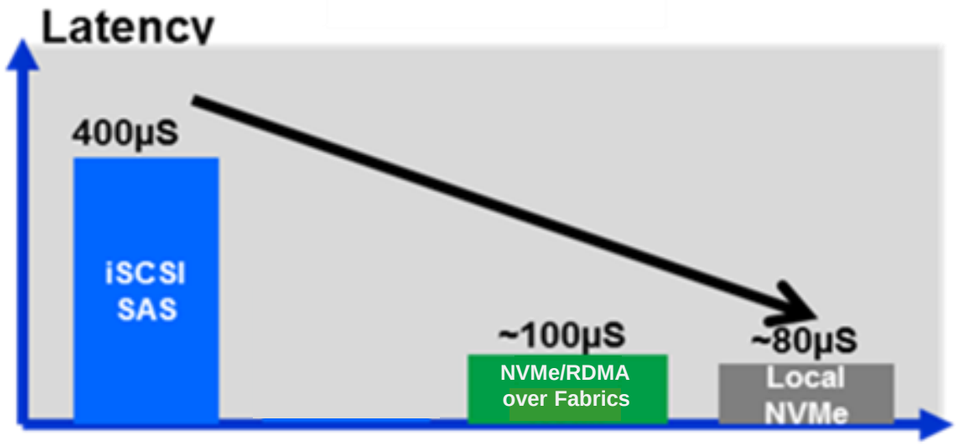
NVMe over Fabrics (NVMe-oF) Transports

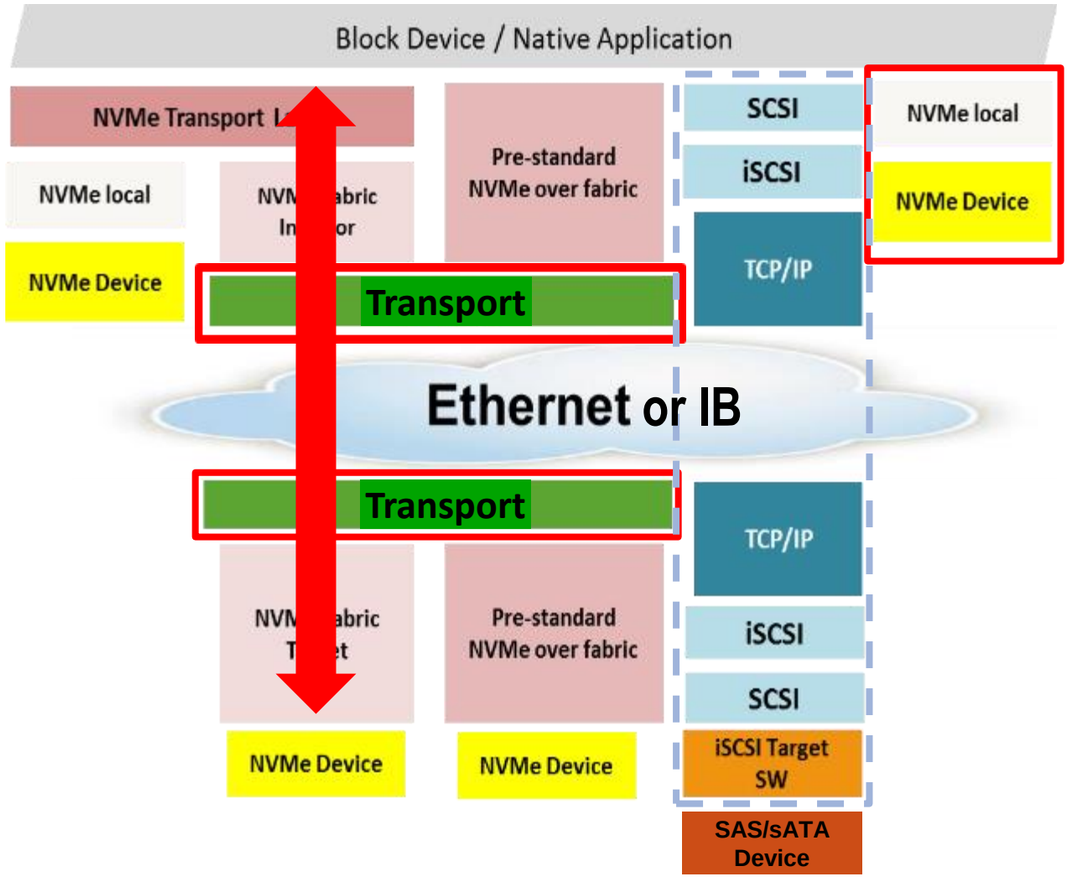
- The NVMe-oF standard is not Fabric specific
- Instead there is a separate Transport Binding specification for each Transport Layer
  - RDMA was 1st
  - Later Fibre Channel
- NVM.org just released a new binding specification for TCP



How Does NVMe-oF Maintain NVMe Performance?

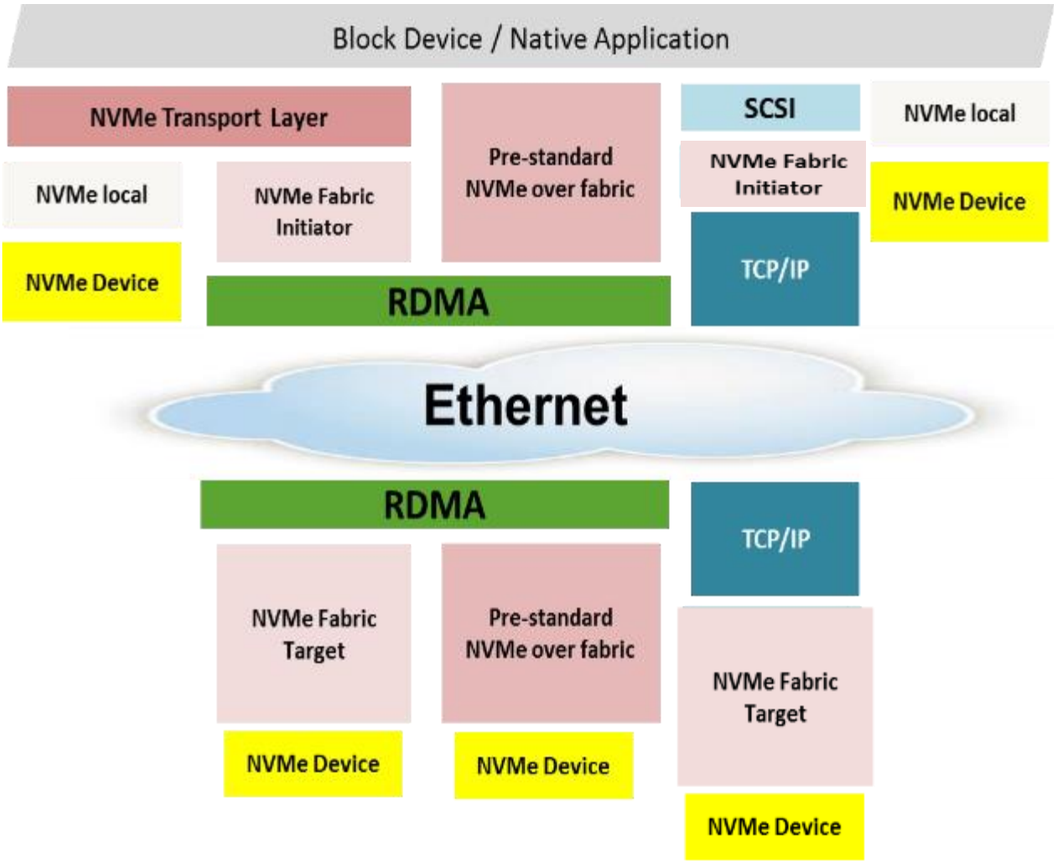
- By extending NVMe efficiency over a fabric
  - NVMe commands and data structures are transferred end to end
- Bypassing legacy stacks for performance
- First products and early demos all used RDMA
- Performance is impressive

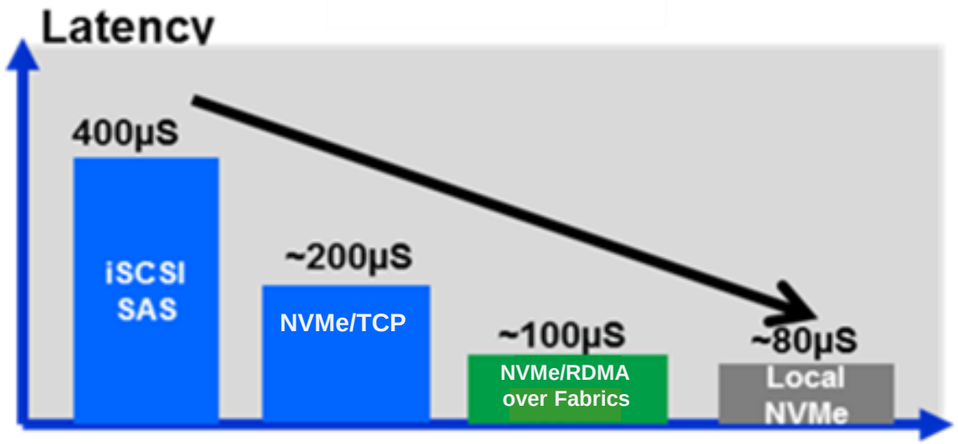




How Does NVMe-oF Maintain NVMe Performance?

- By extending NVMe efficiency over a fabric
  - NVMe commands and data structures are transferred end to end
- Bypassing legacy stacks for performance
- First products and early demos all used RDMA
- Performance is impressive

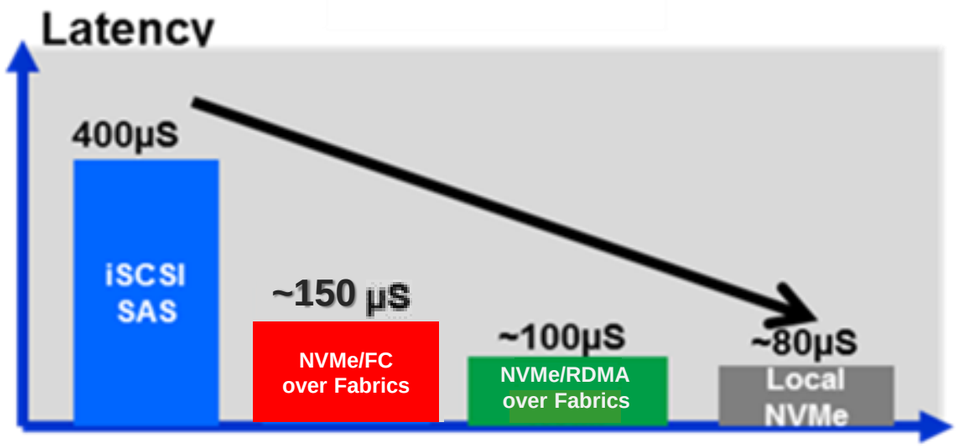


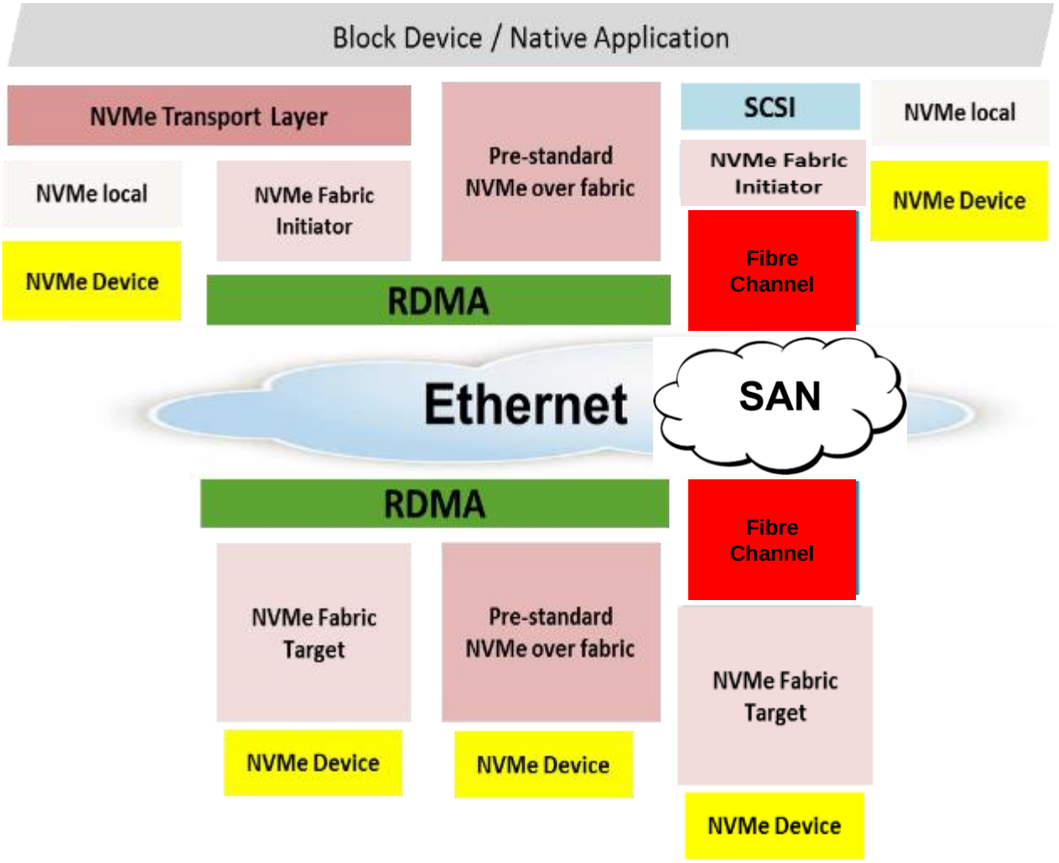


https://www.theregister.co.uk/2018/08/16/pavilion_fabrics_performance/

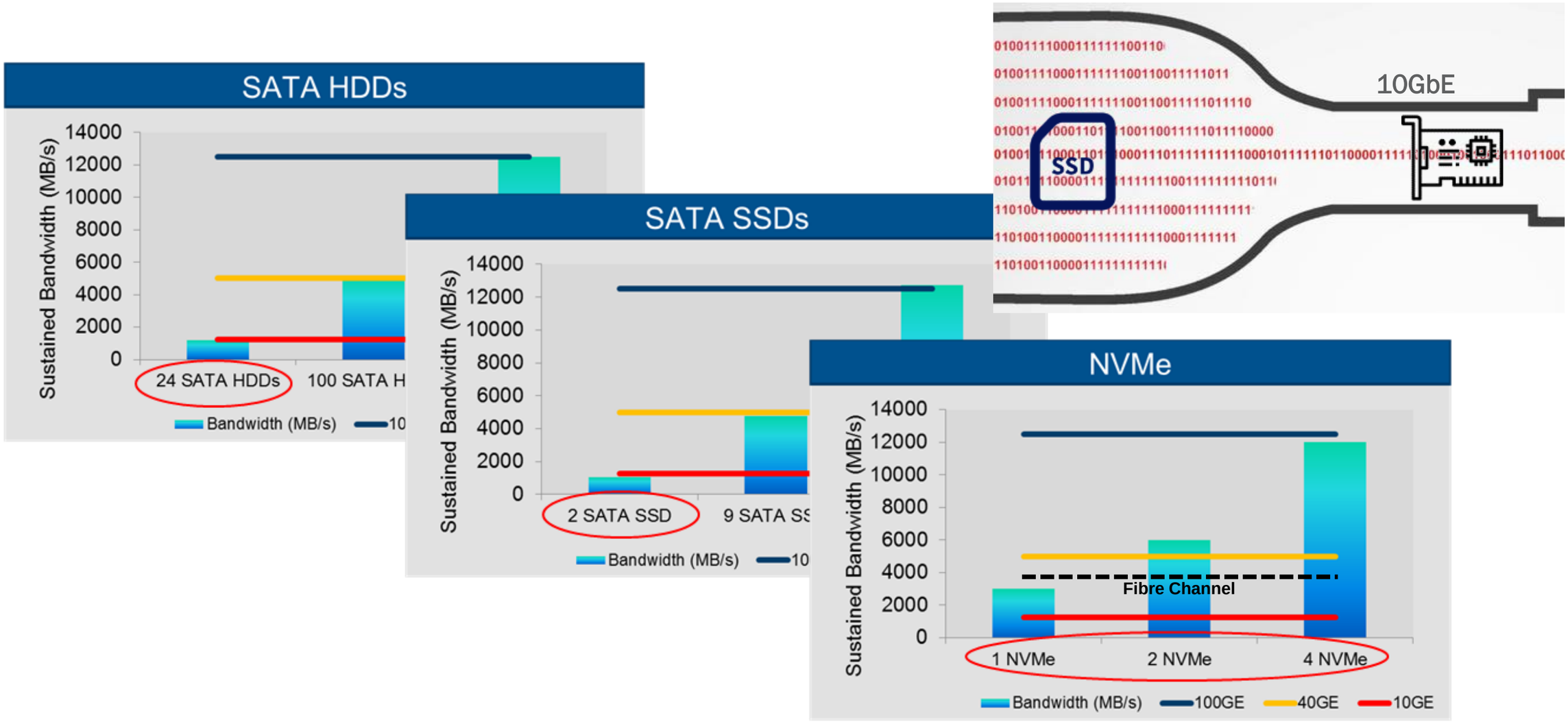
How Does NVMe-oF Maintain NVMe Performance?

- By extending NVMe efficiency over a fabric
  - NVMe commands and data structures are transferred end to end
- Bypassing legacy stacks for performance
- First products and early demos all used RDMA
- Performance is impressive

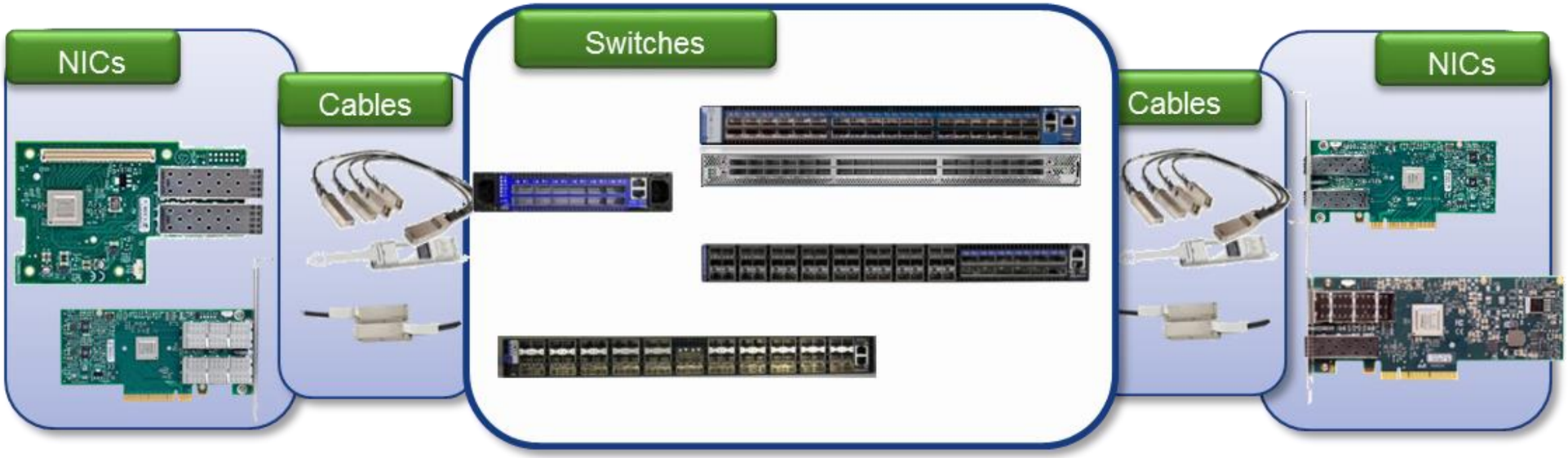




Faster Storage Needs a Faster Network

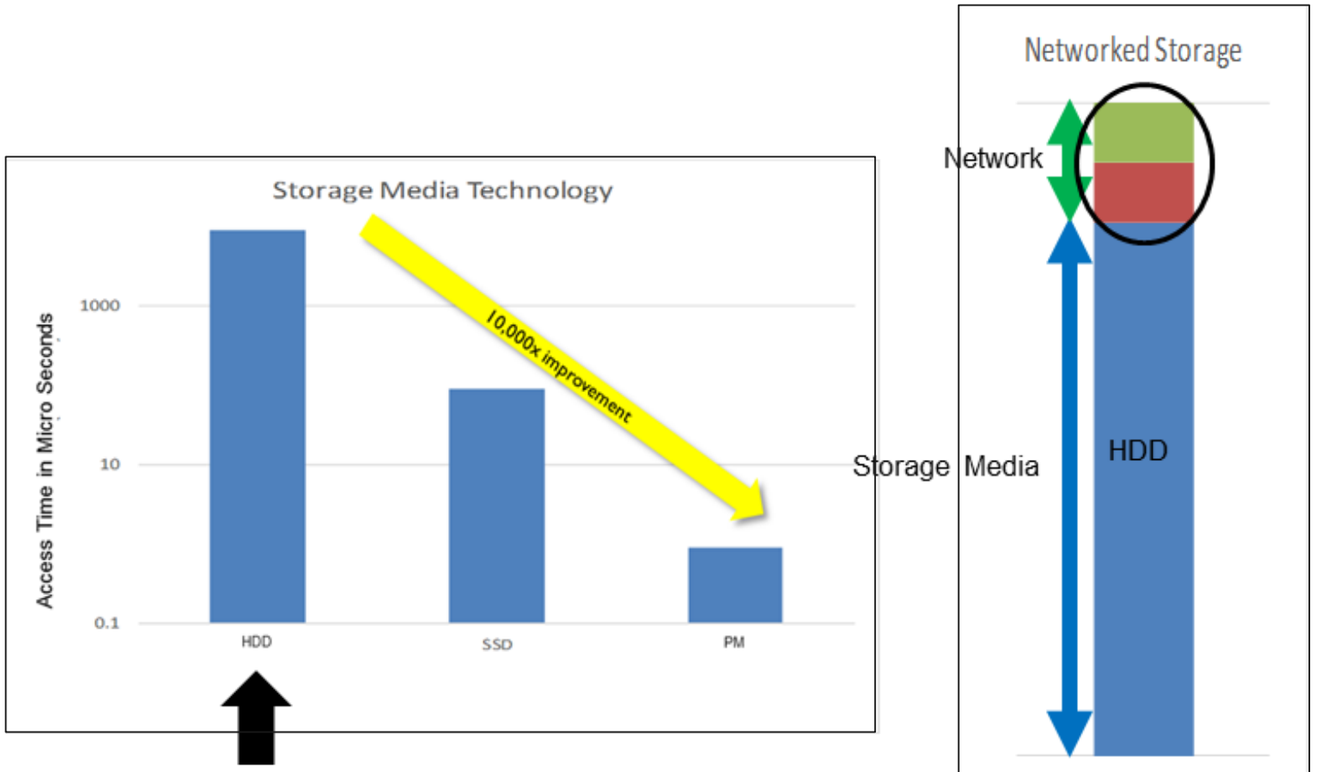


Faster Network Wires Solves Some the Network Bottle Neck Problem…

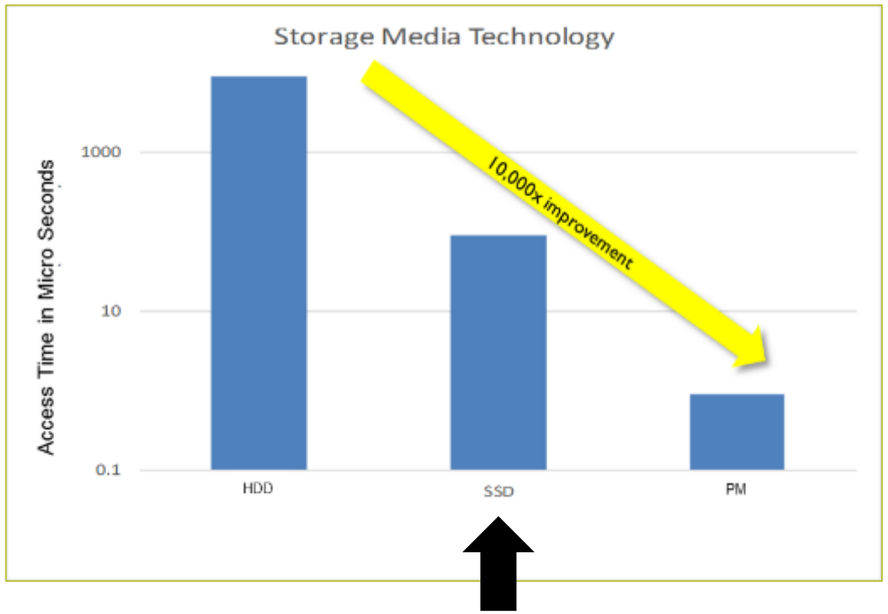


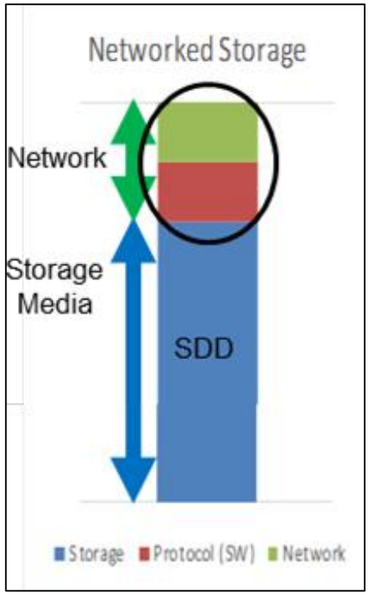
Ethernet & InfiniBand
End-to-End 25, 40, 50, 56, 100, 200Gb
Going to 400Gb

Faster Protocols Solves the Rest

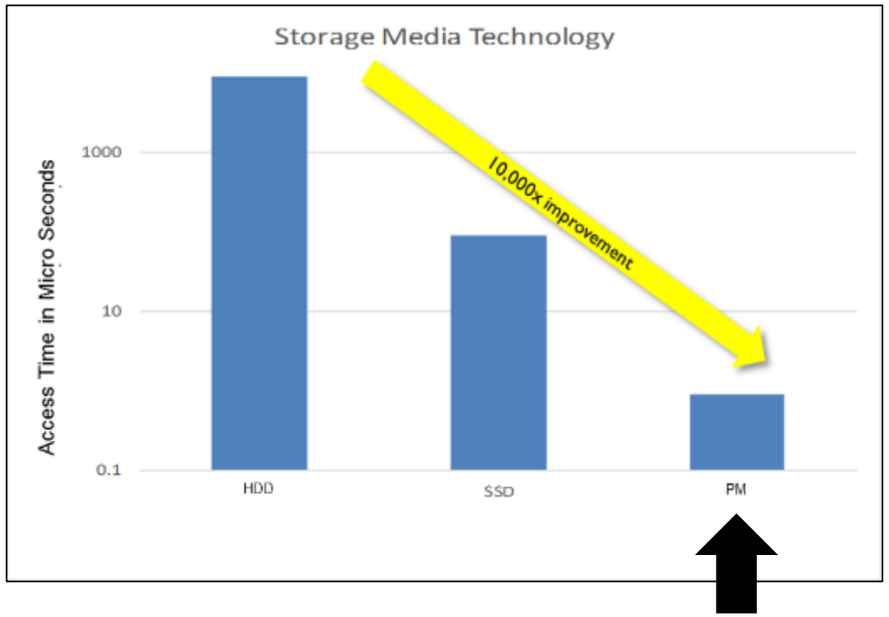


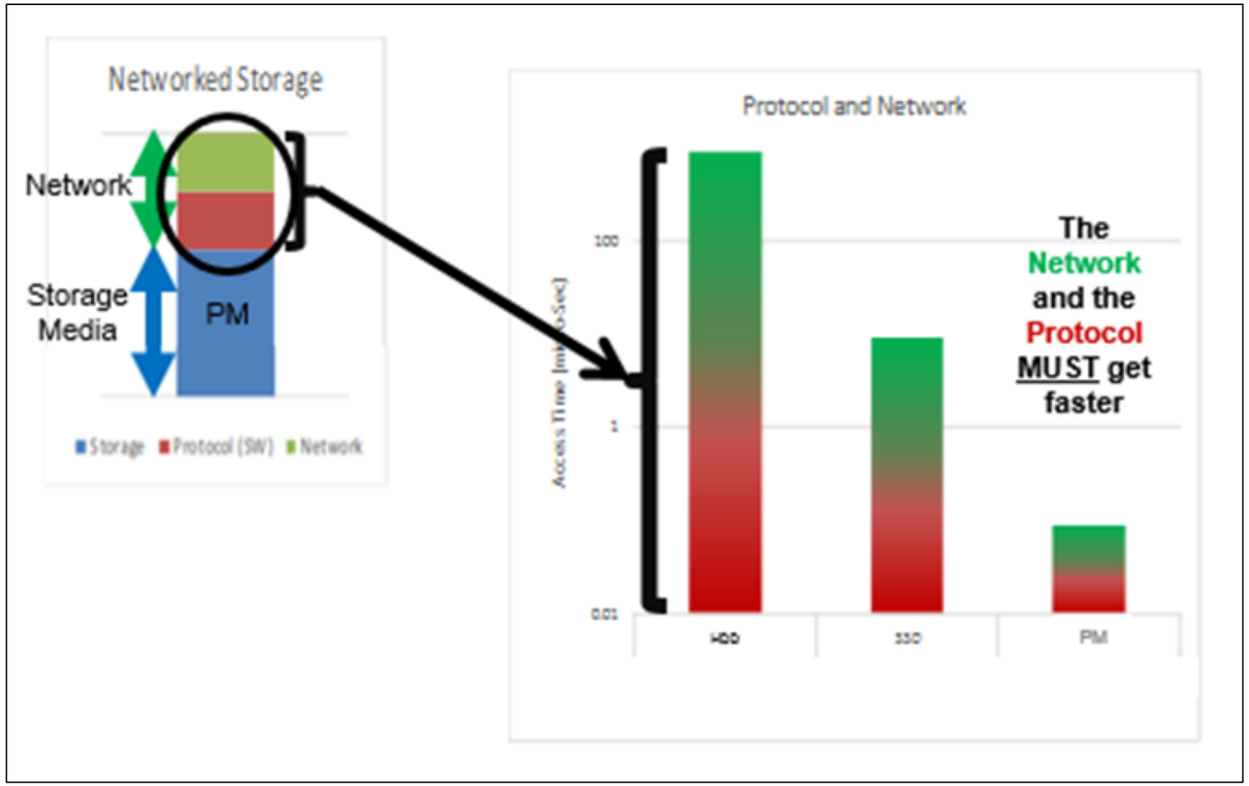
Faster Protocols Solves the Rest





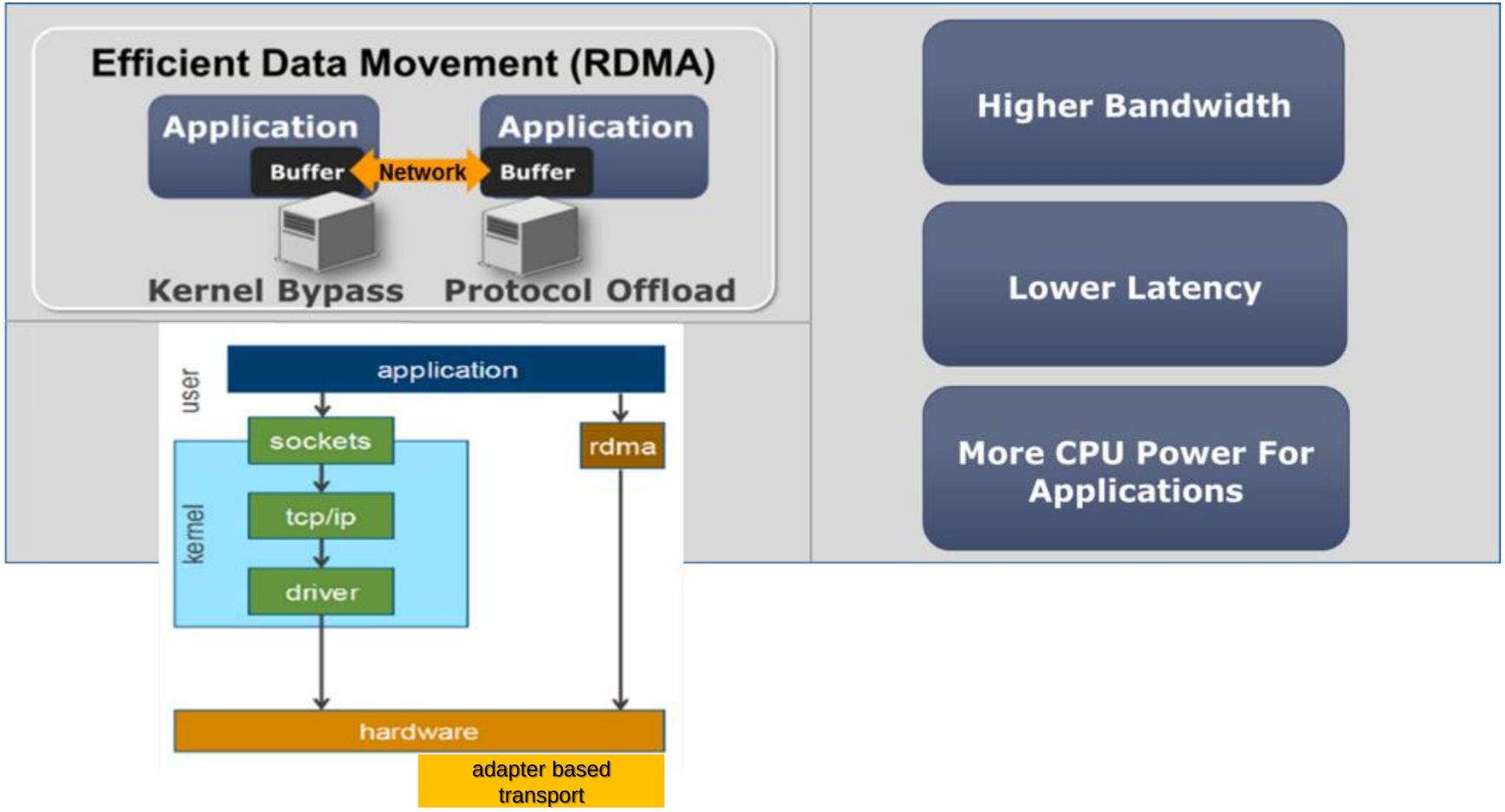
NVMe, NVMe-oF, and RDMA Protocols



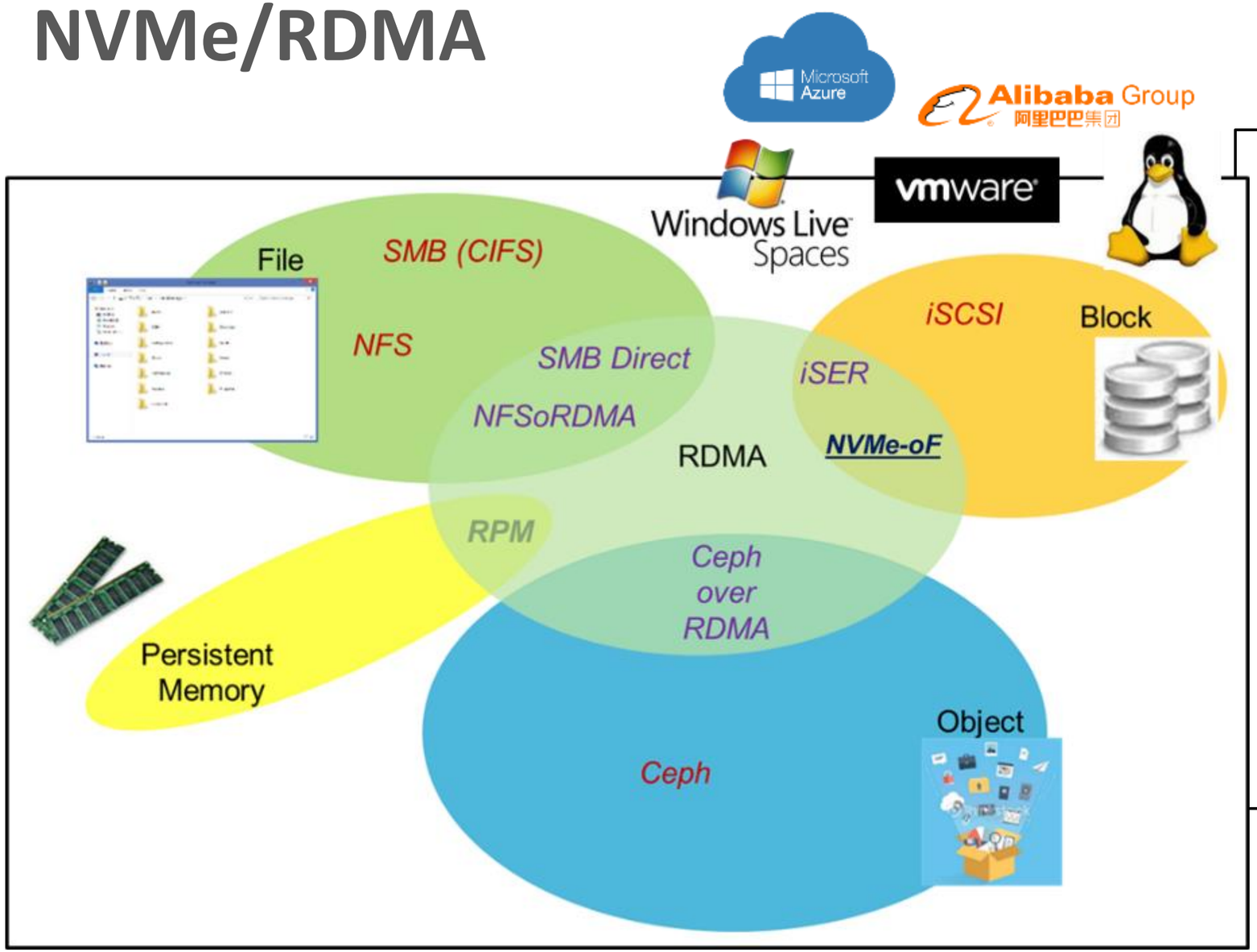


NVMe/RDMA

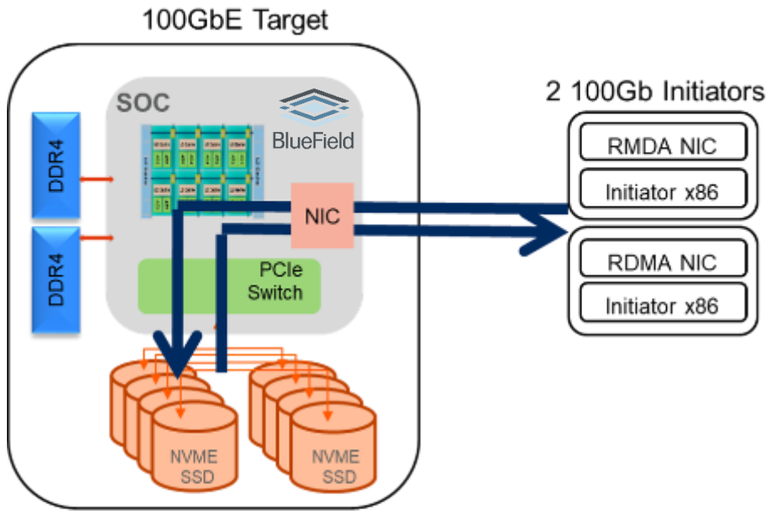
NVMe-oF over RoCE



NVMe/RDMA

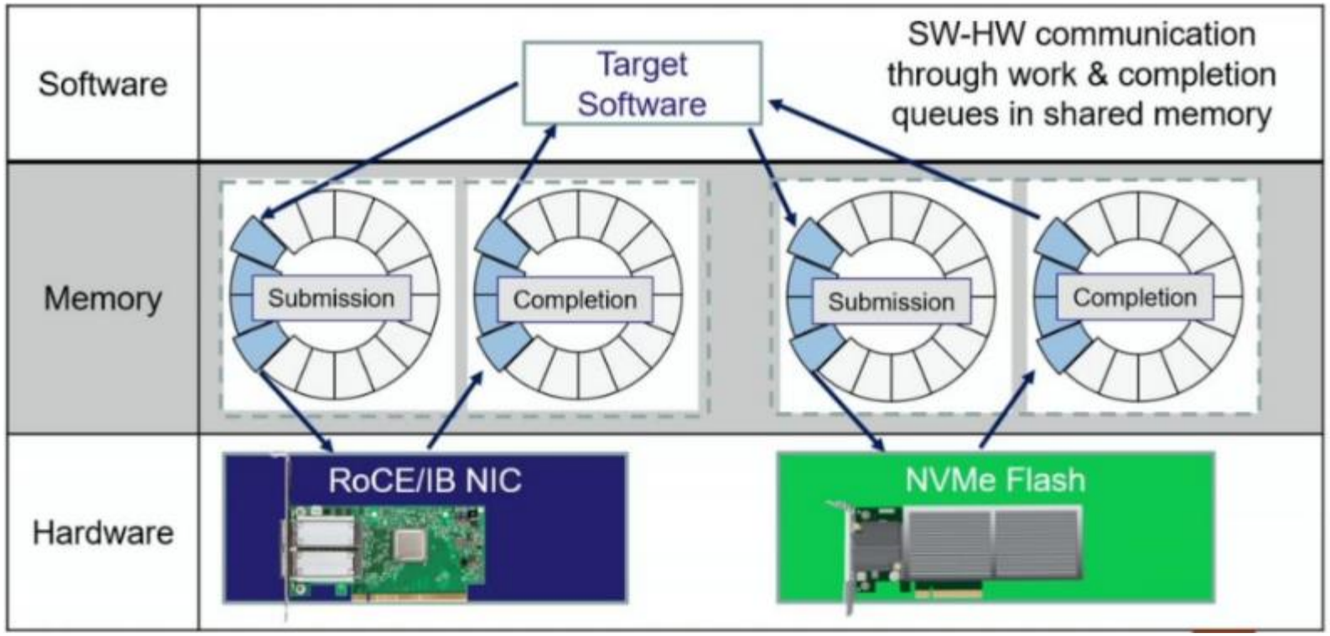


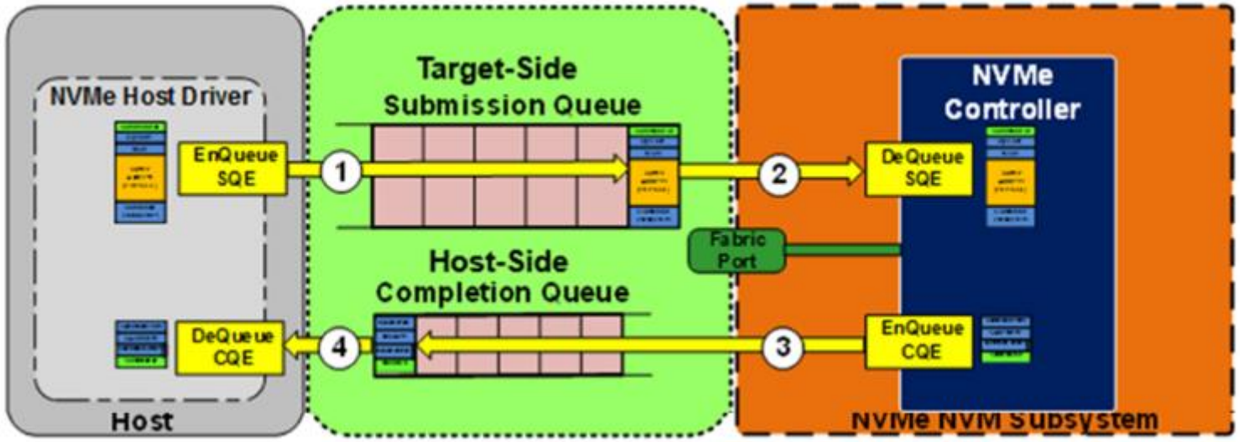
NVMe-oF over RoCE

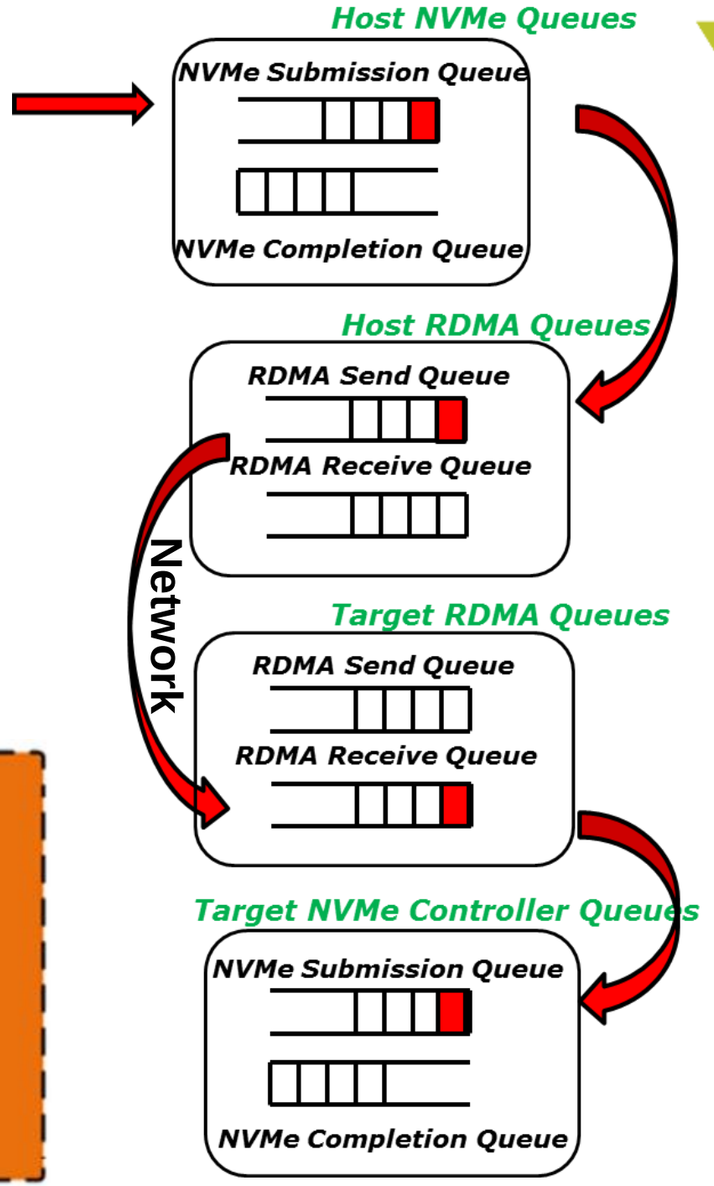


- 8M IOPs, 512B block size
- 5M IOPs, 4K block side
- 0.01% CPU utilization
- ~3usec latency

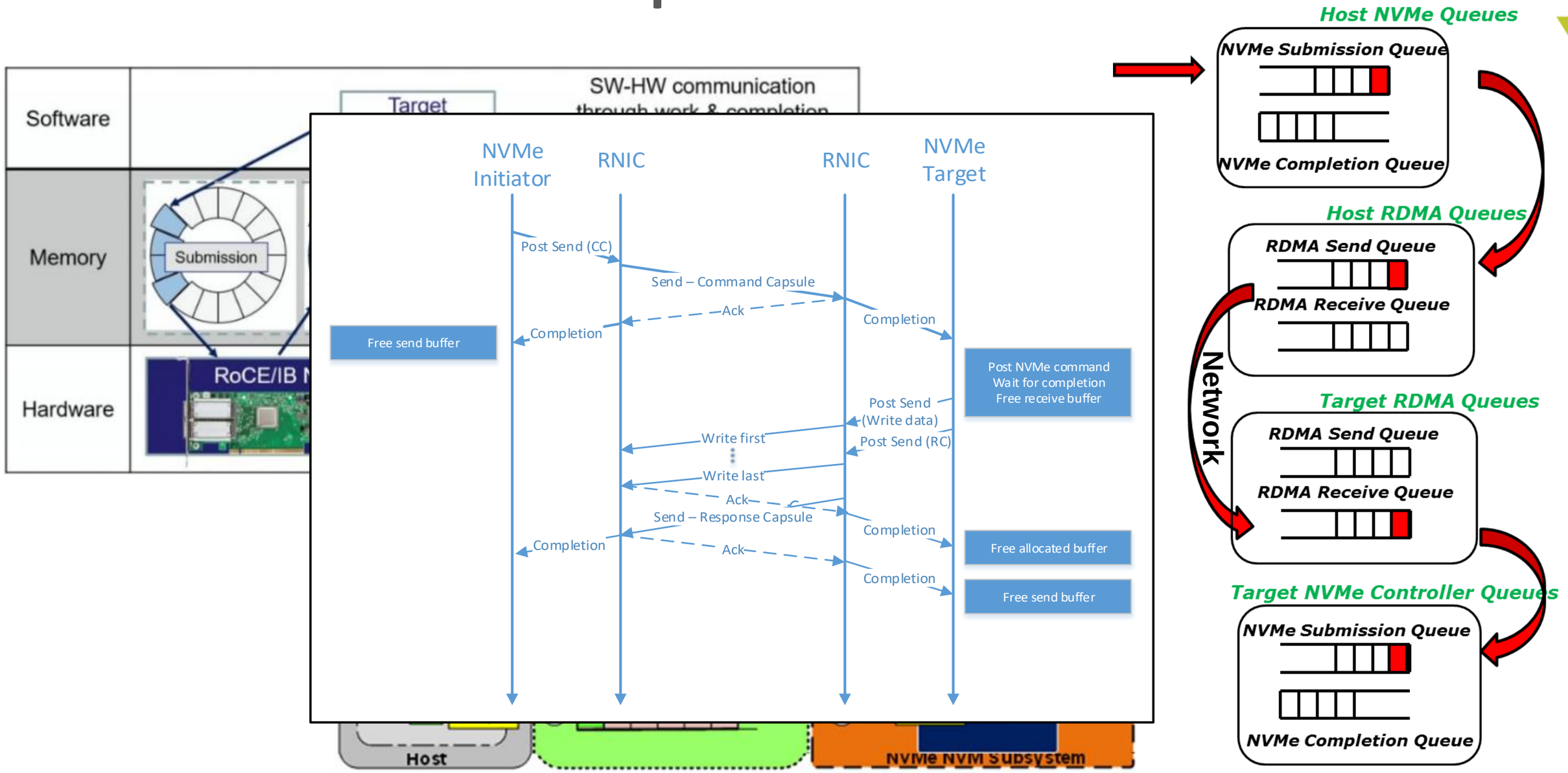
NVMe Commands Encapsulated



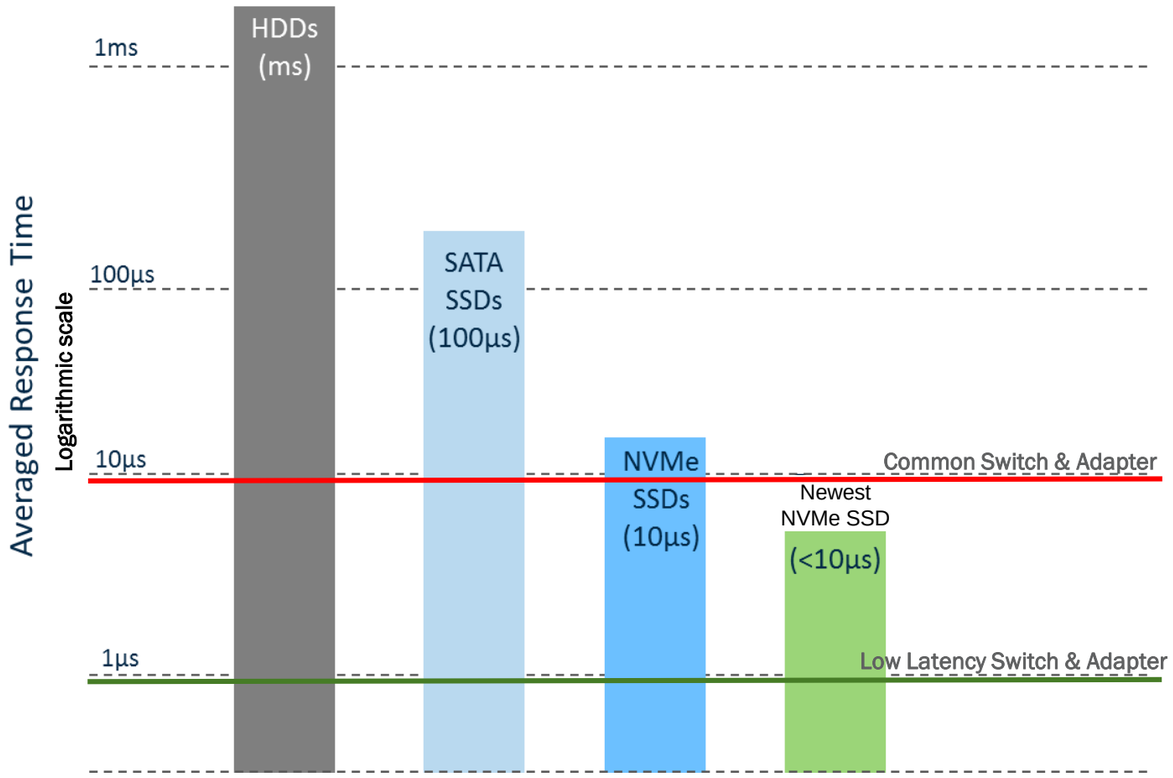


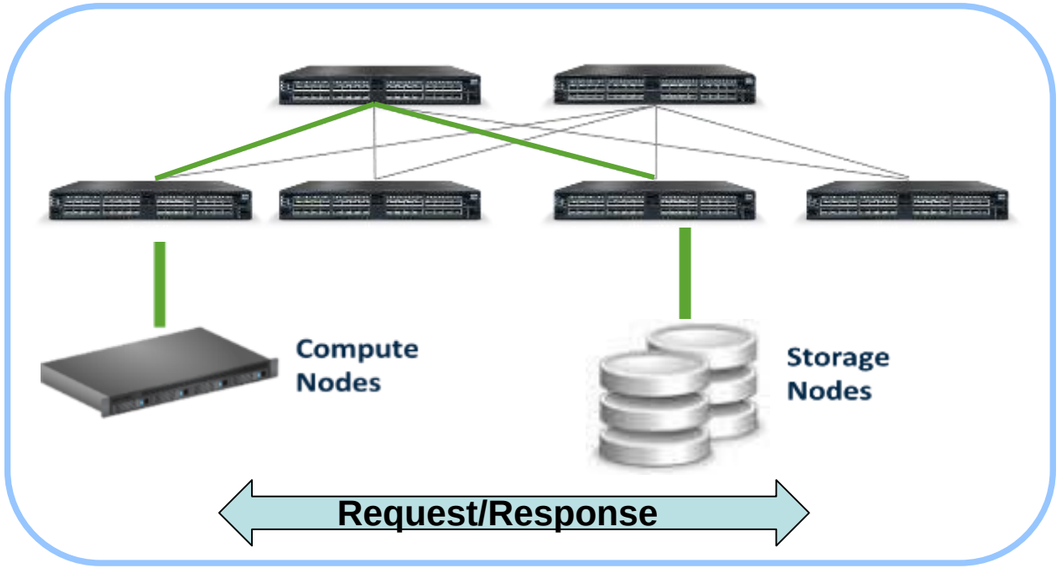


NVMe Commands Encapsulated



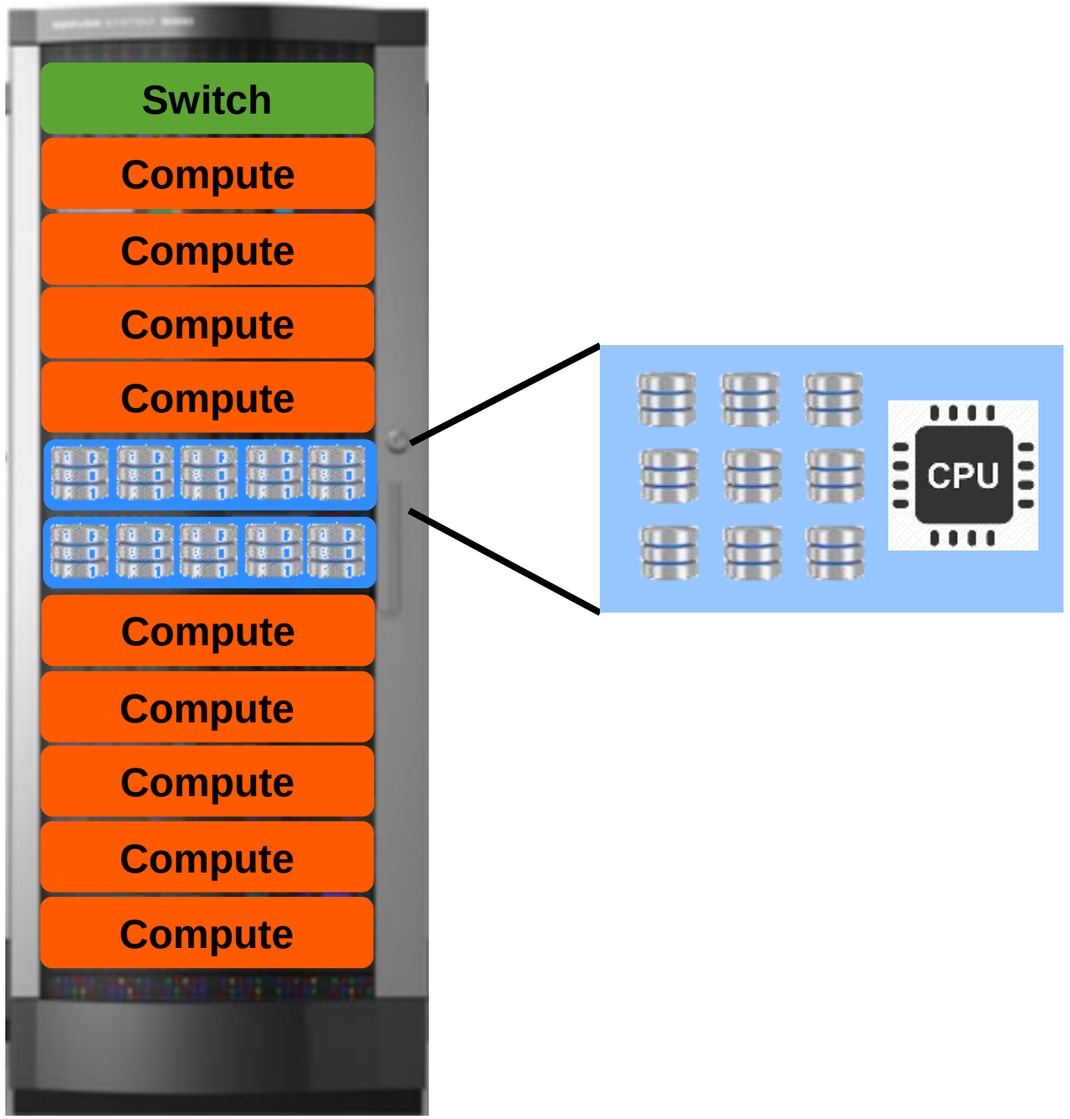
Importance of Latency with NVMe-oF





Network hops multiply latency

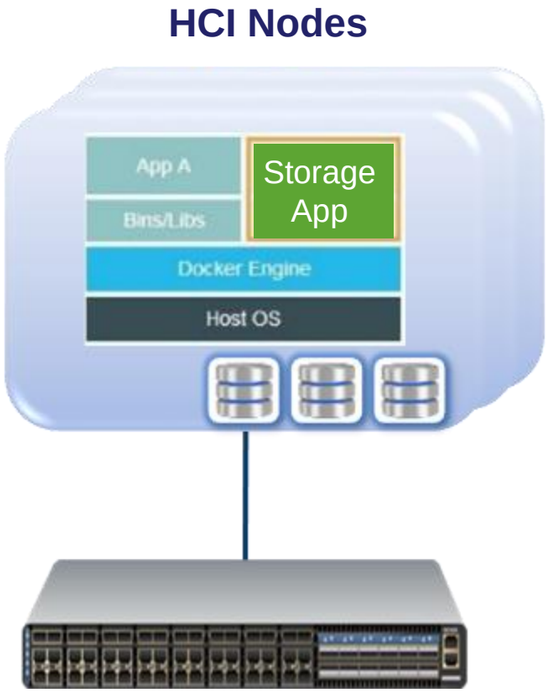
Composable Infrastructure Use Case



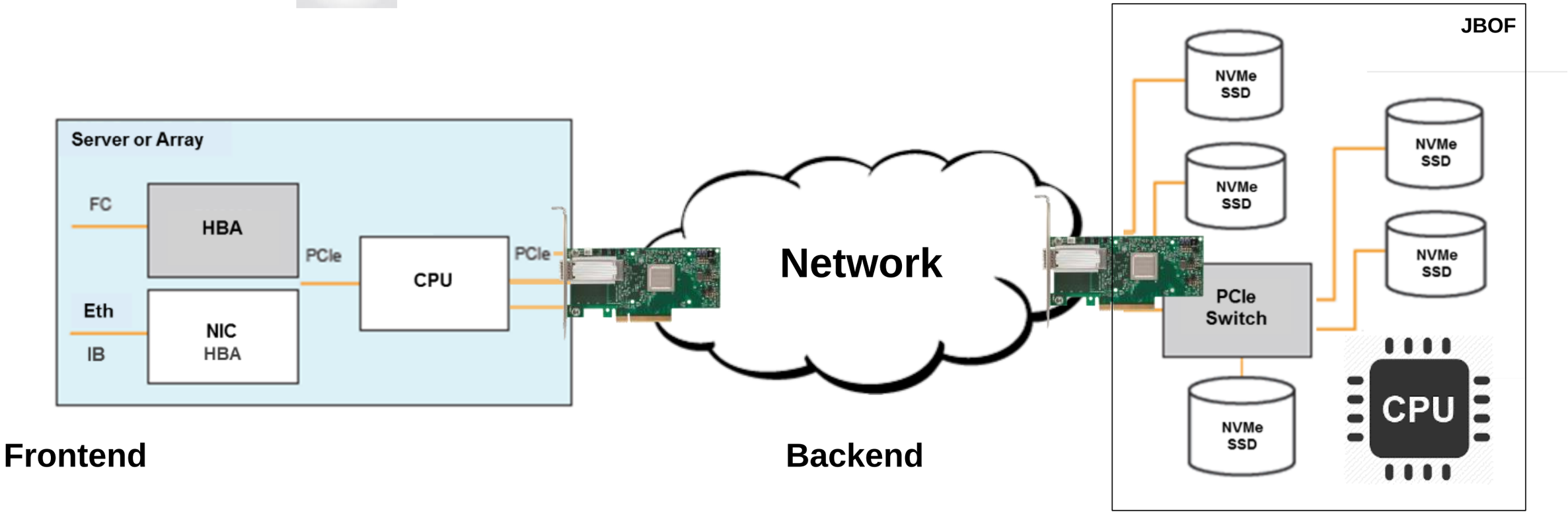
- Also called Compute Storage Disaggregation and Rack Scale
  - Dramatically improves data center efficiency
- NVMe over Fabrics enables Composable Infrastructure
  - Low latency
  - High bandwidth
  - Nearly local disk performance

Hyperconverged and Scale-Out Storage Use Case

- Scale-out
  - Cluster of commodity servers
  - Software provides storage functions
- Hyperconverged collapses compute & storage
  - Integrated compute-storage nodes & software
  - NVMe-oF performs like local/direct-attached SSD

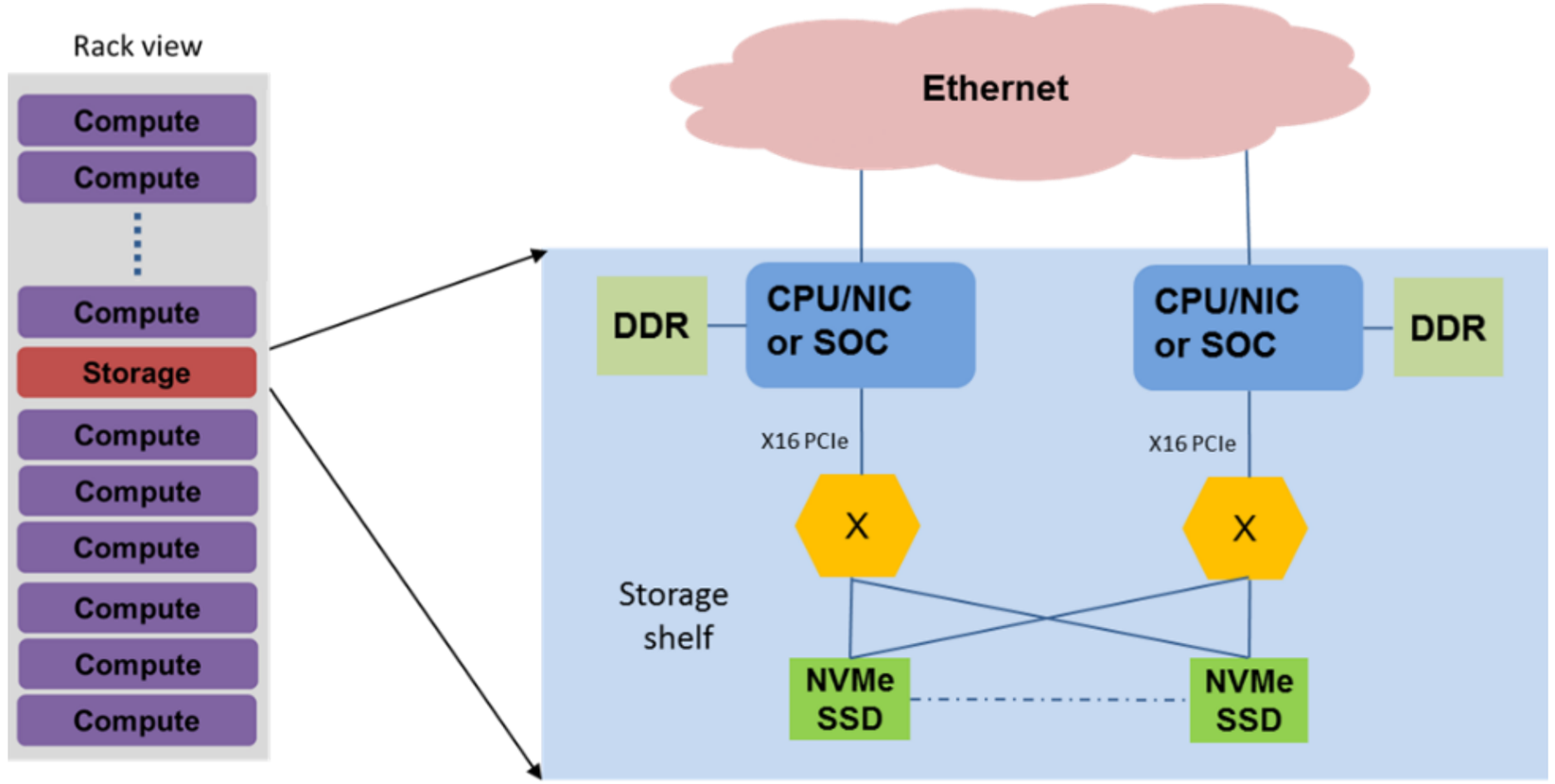


Backend Scale Out Use Case

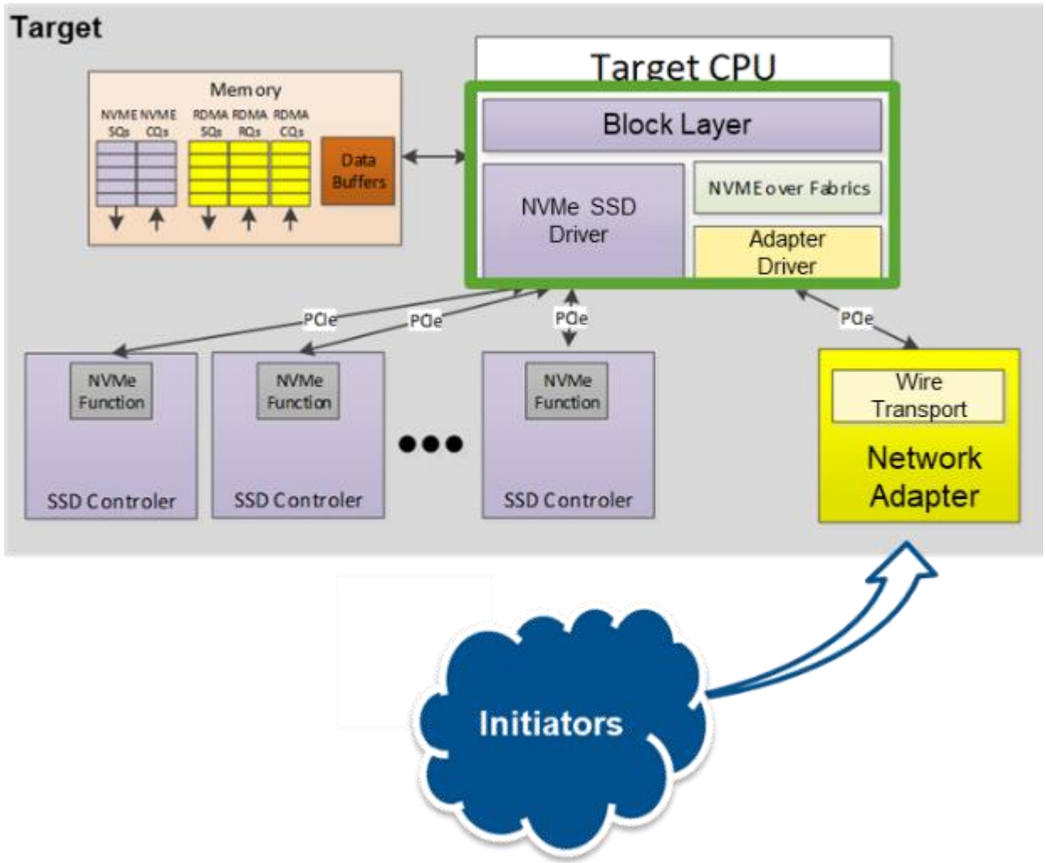


NVMe-oF Use Cases: Classic SAN

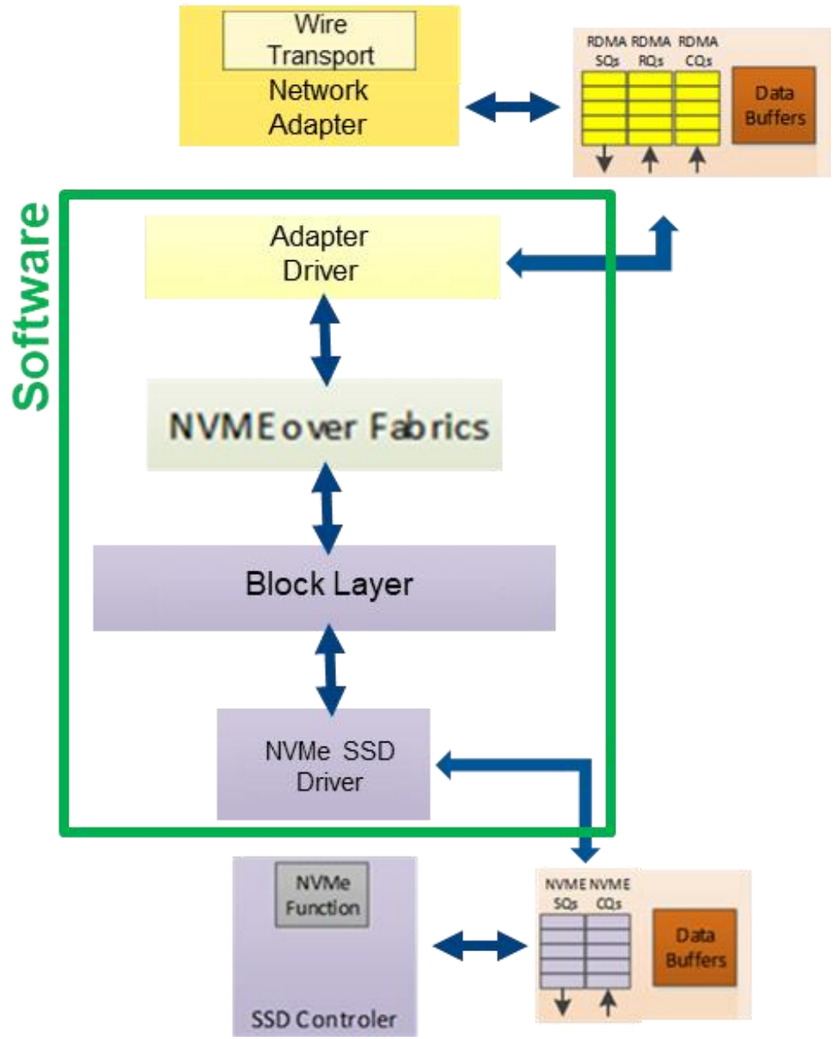
- SAN features at higher performance
  - Better utilization: capacity, rack space, and power
  - Scalability
  - Management
  - Fault isolation



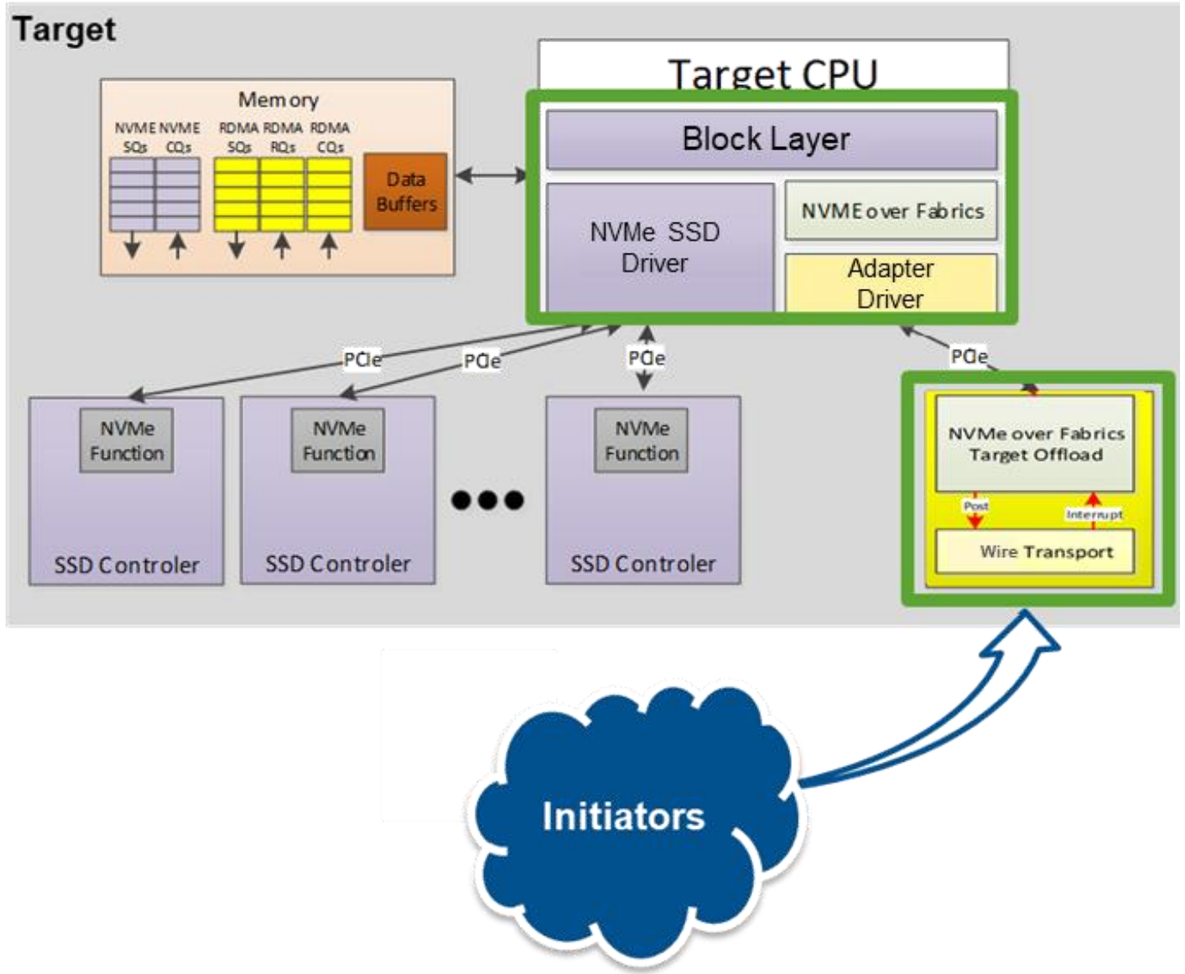
NVMe-oF Target Hardware Offloads



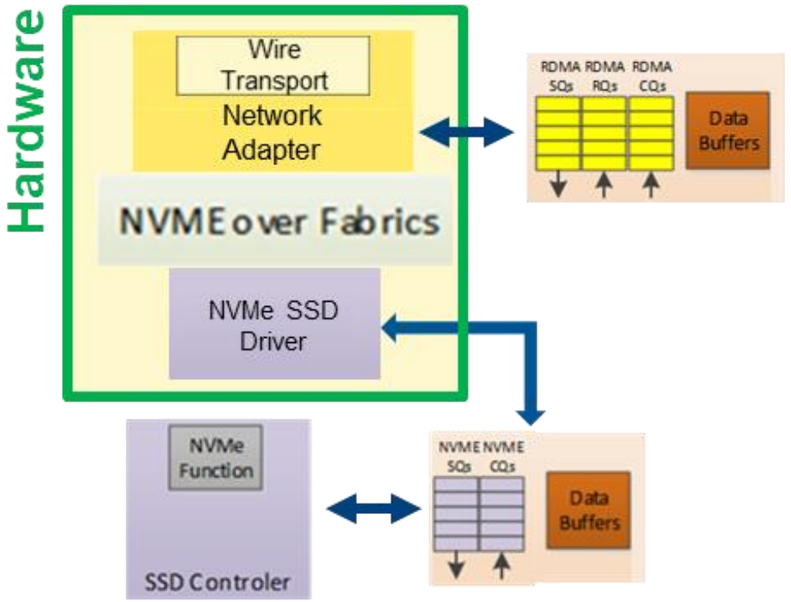
No Offload Mode



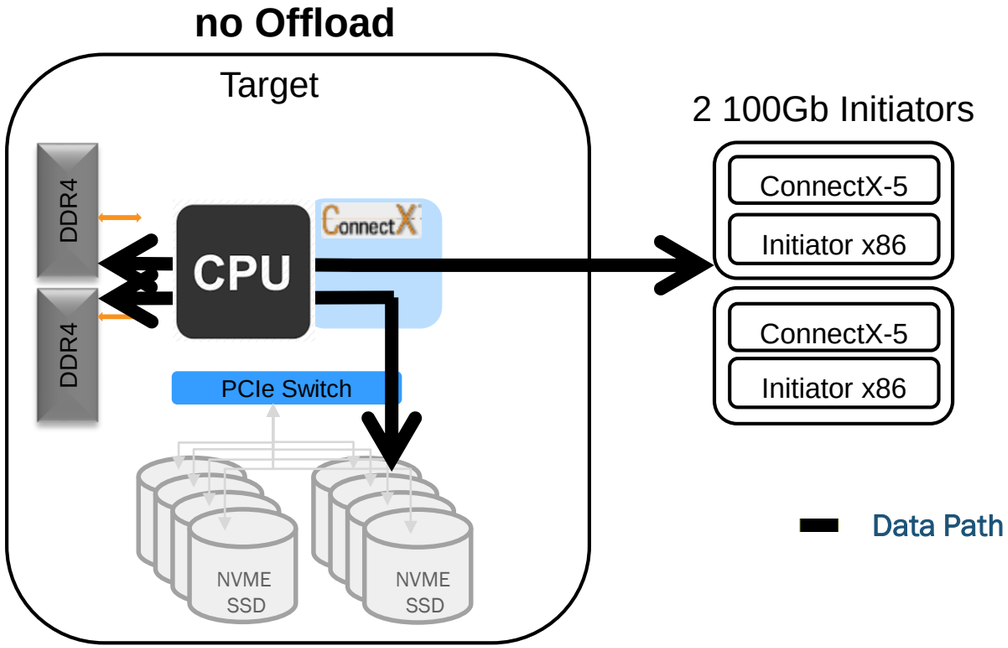
How Target Offload Works

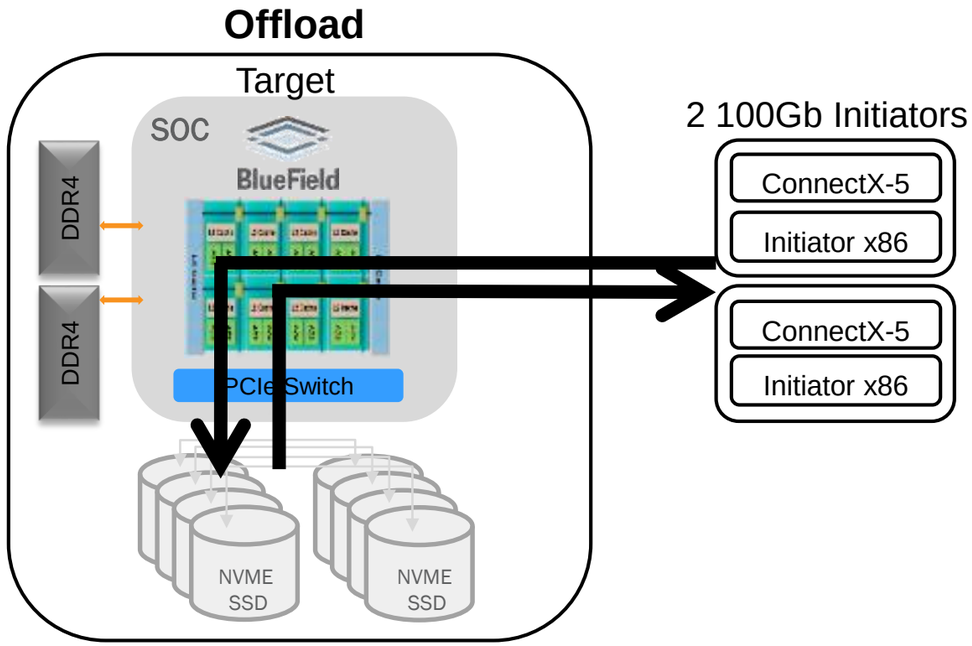


- Offload
  - Only control path, management and exceptions go through Target CPU software
  - Data path and NVMe commands handled by the network adapter



Offload vs No Offload Performance

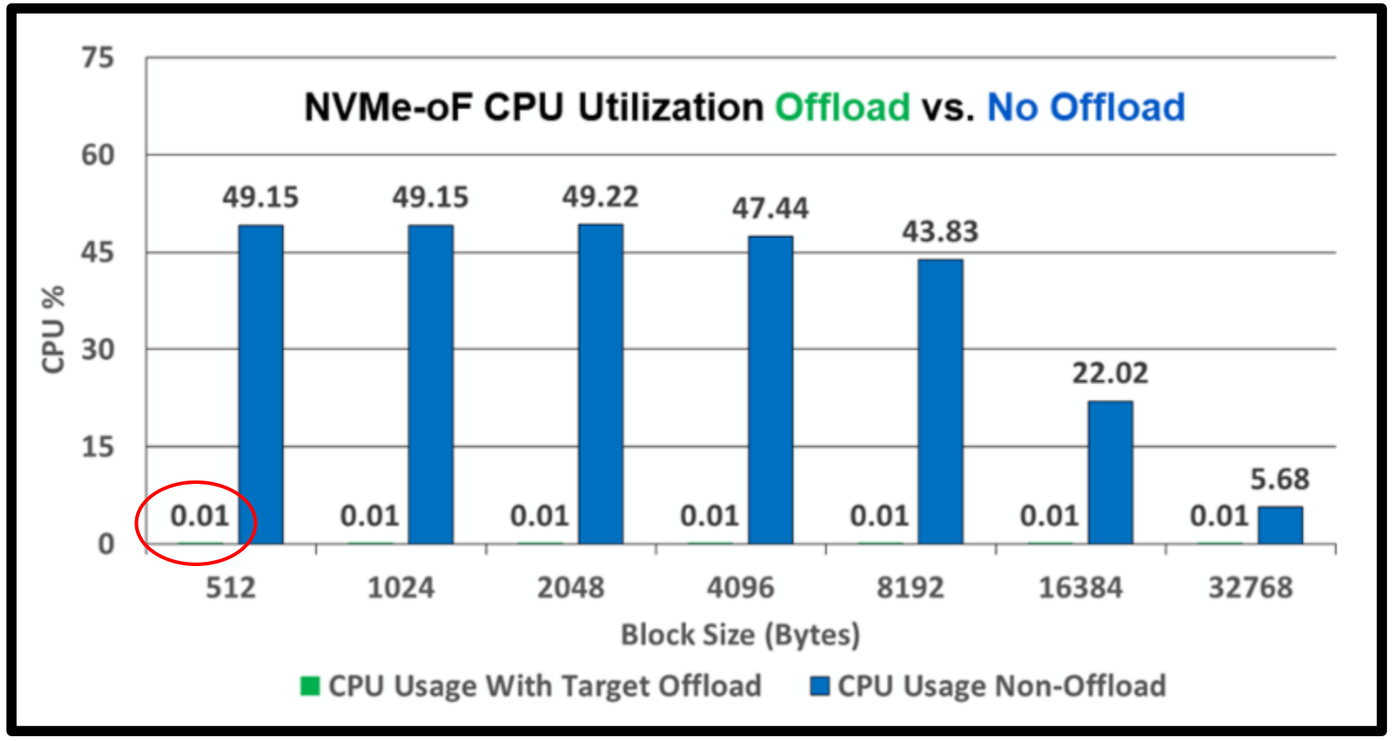


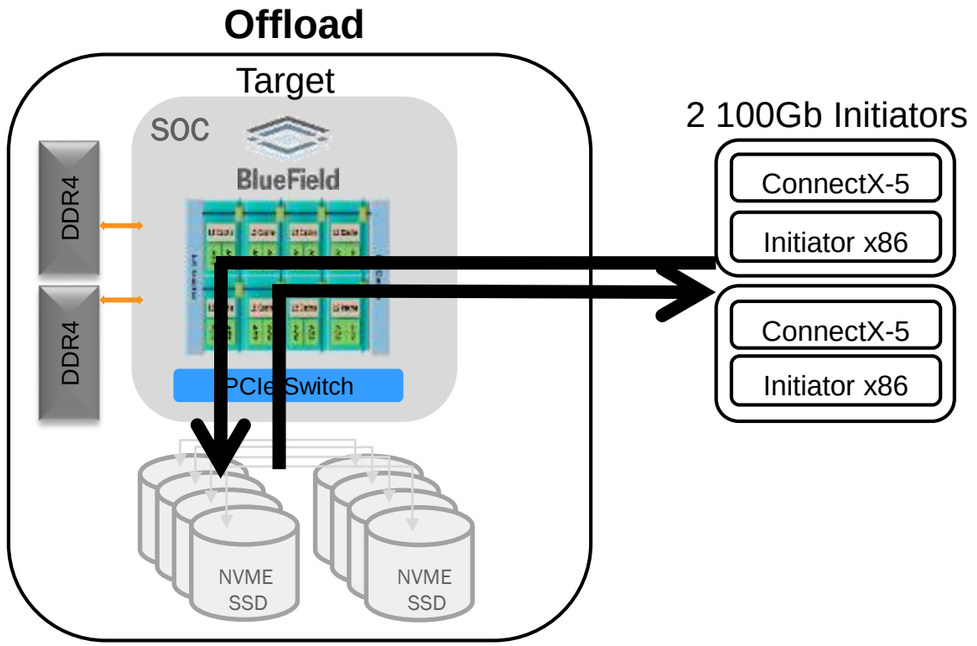


- 6M IOPs, 512B block size
- 2M IOPs, 4K block side
- ~15 usec latency (not including SSD)

- 8M IOPs, 512B block size
- 5M IOPs, 4K block side
- ~5 usec latency (not including SSD)

Offload vs No Offload Performance



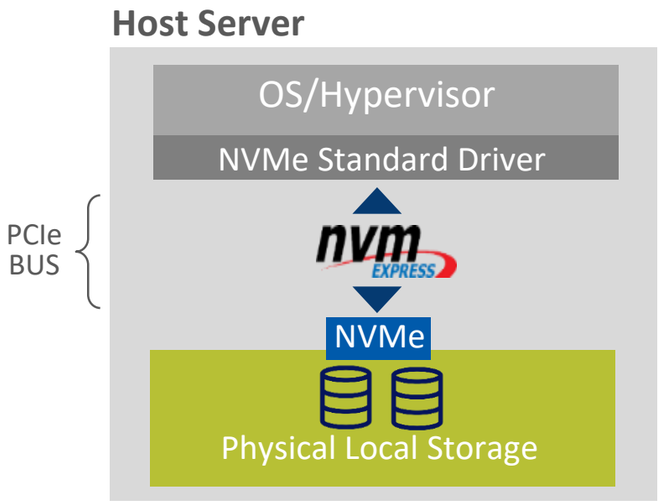


- 6M IOPs, 512B block size
- 2M IOPs, 4K block side
- ~15 usec latency (not including SSD)

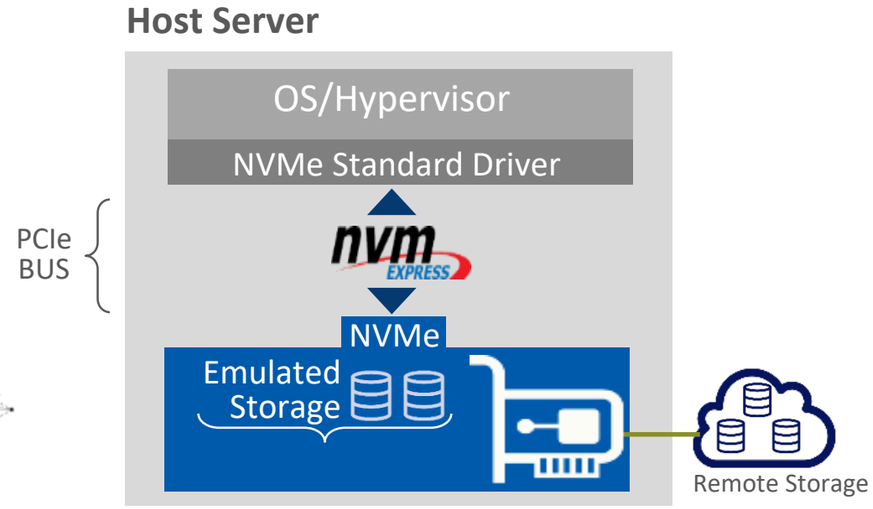
- 8M IOPs, 512B block size
- 5M IOPs, 4K block side
- ~5 usec latency (not including SSD)

NVMe Emulation

Physical Local NVMe Storage

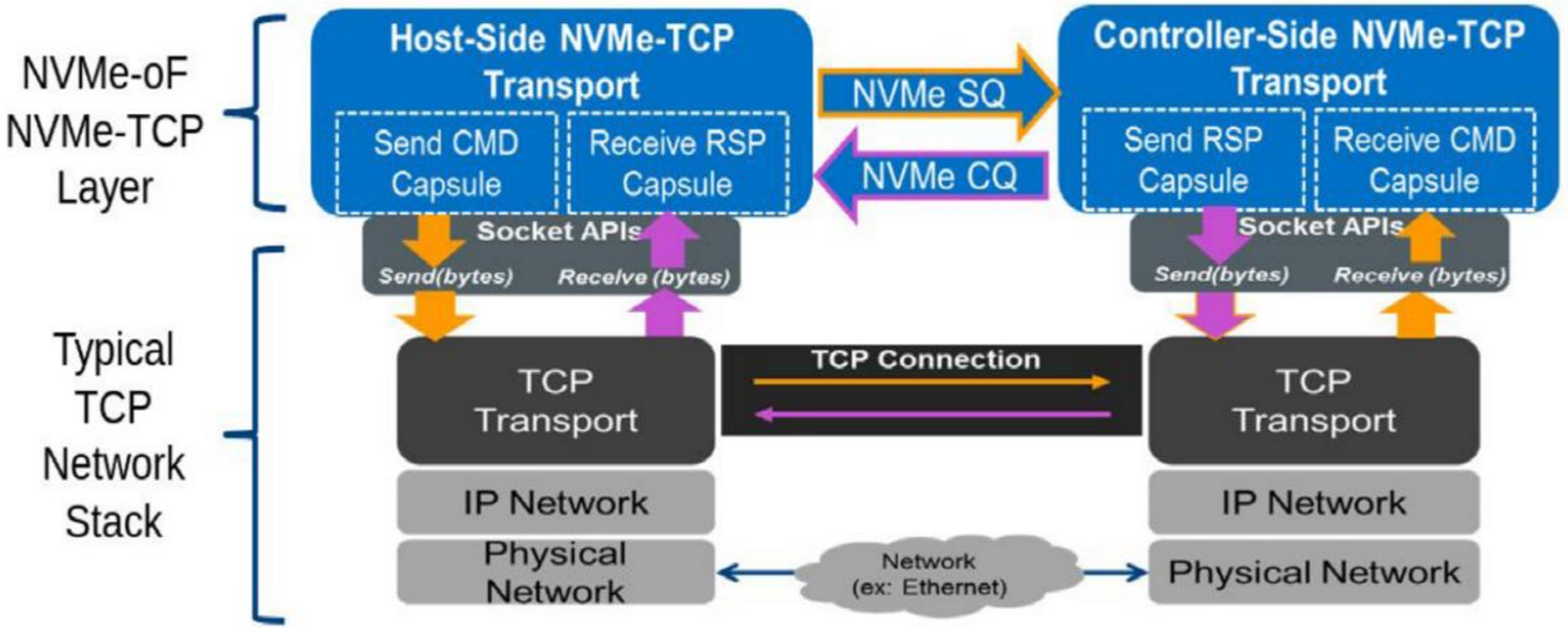


NVMe Drive Emulation



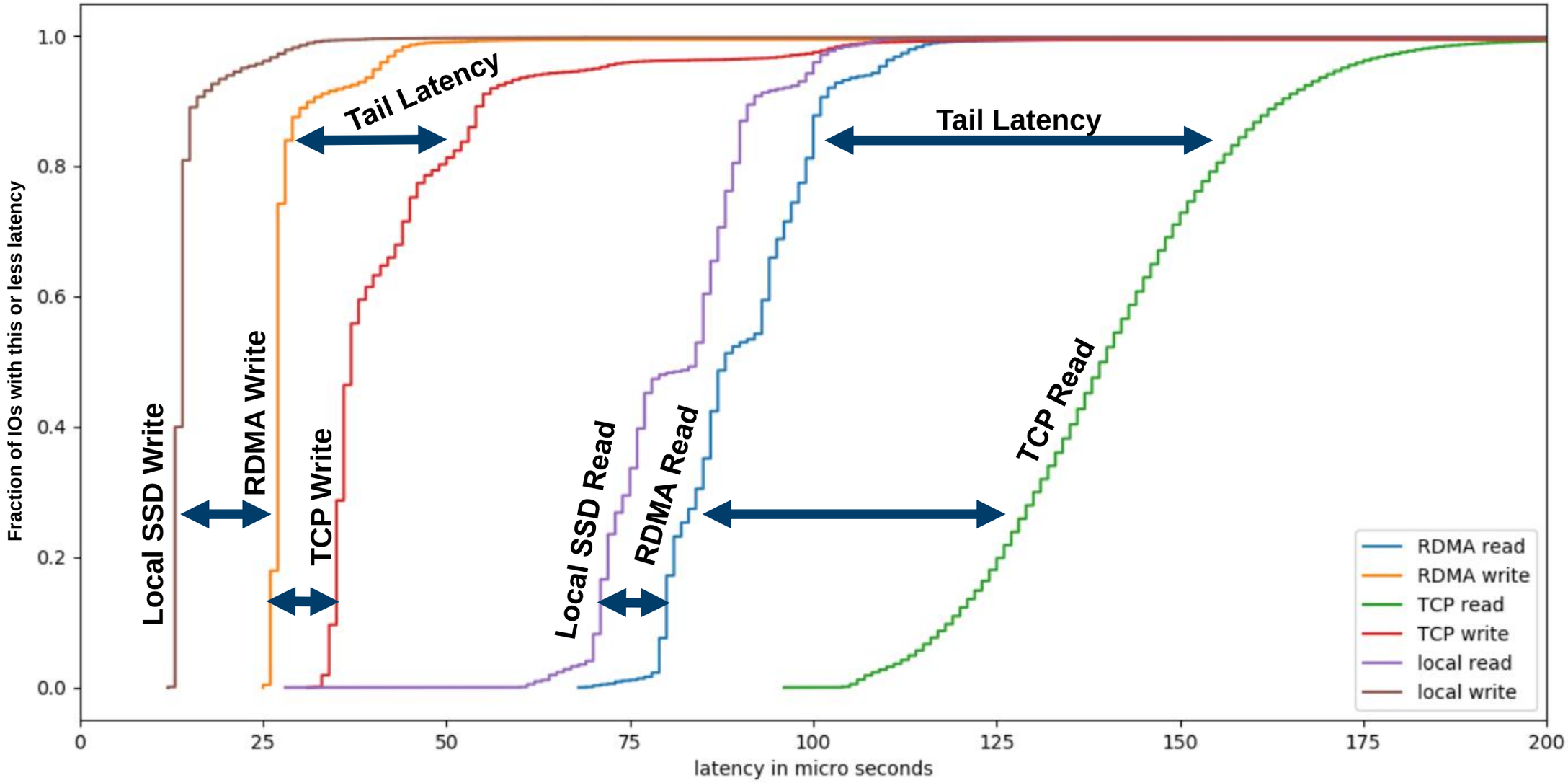
Local Physical Storage to Hardware Emulated Storage

NVMe/TCP

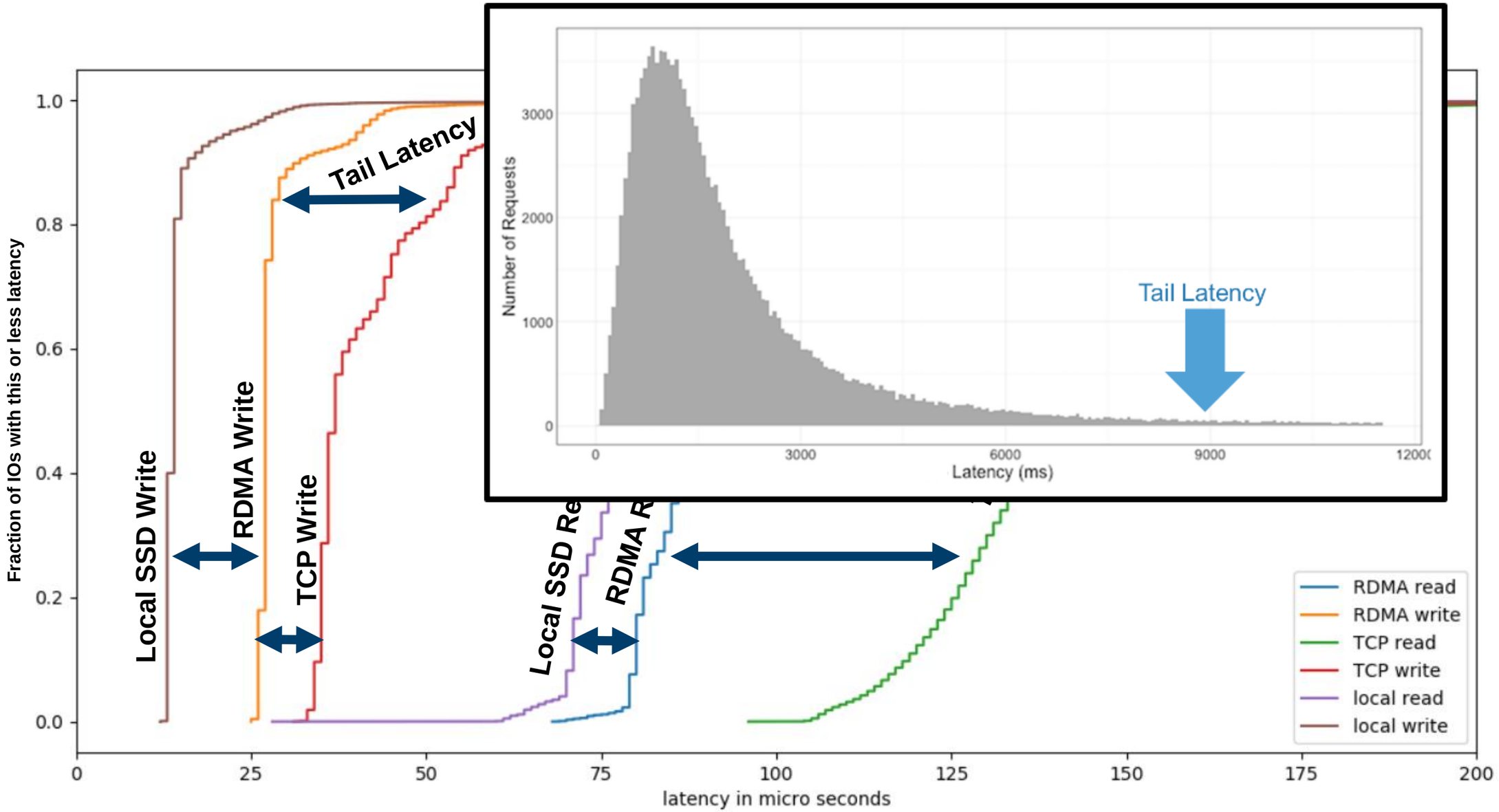


- NVMe-oF commands are sent over standard TCP/IP sockets
- Each NVMe queue pair is mapped to a TCP connection
- Easy to support NVMe over TCP with **no changes**
- Good for distance, stranded server, and out of band management connectivity

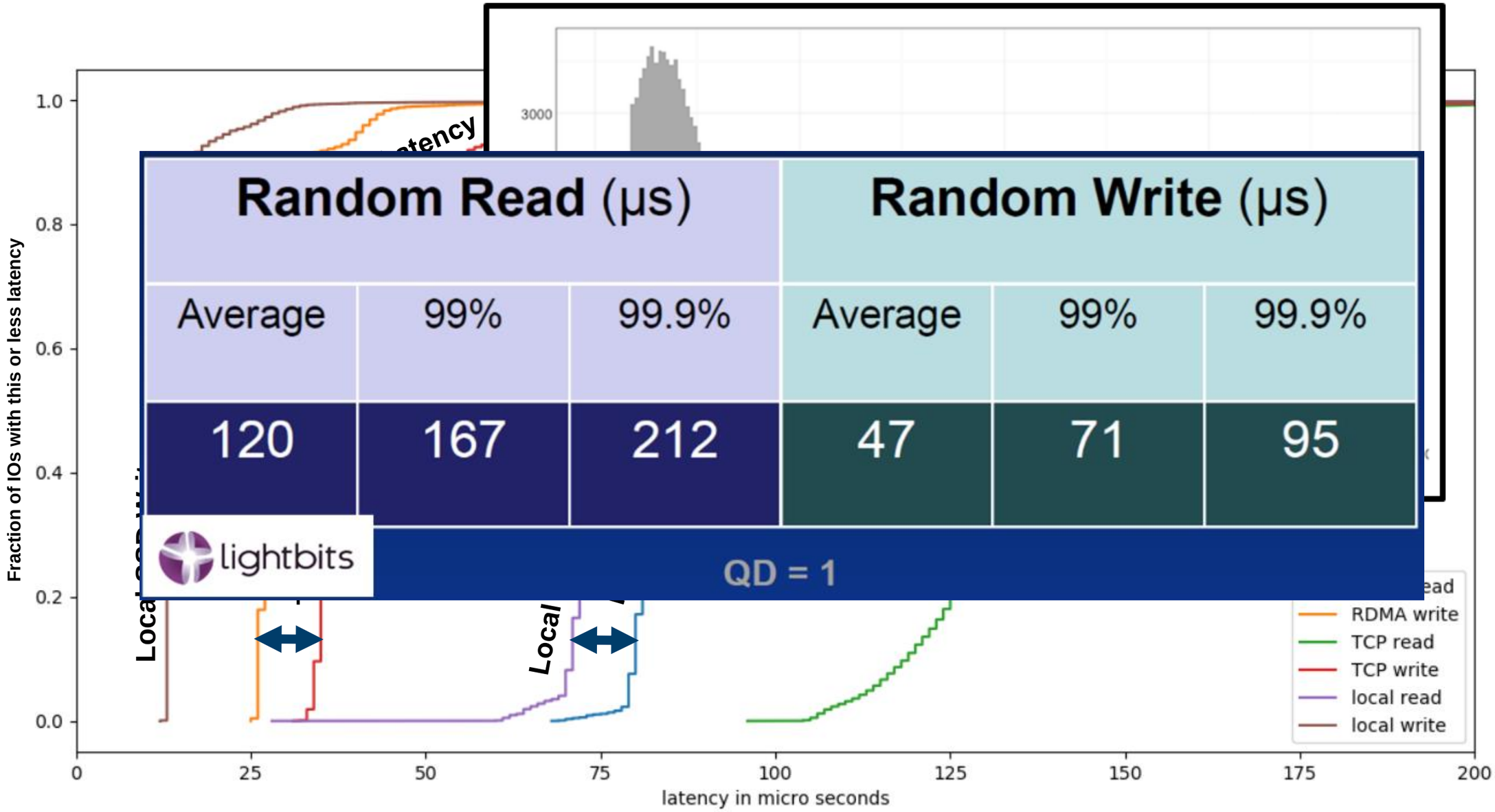
Latency: NVMe-RDMA vs NVMe-TCP



Latency: NVMe-RDMA vs NVMe-TCP



Latency: NVMe-RDMA vs NVMe-TCP

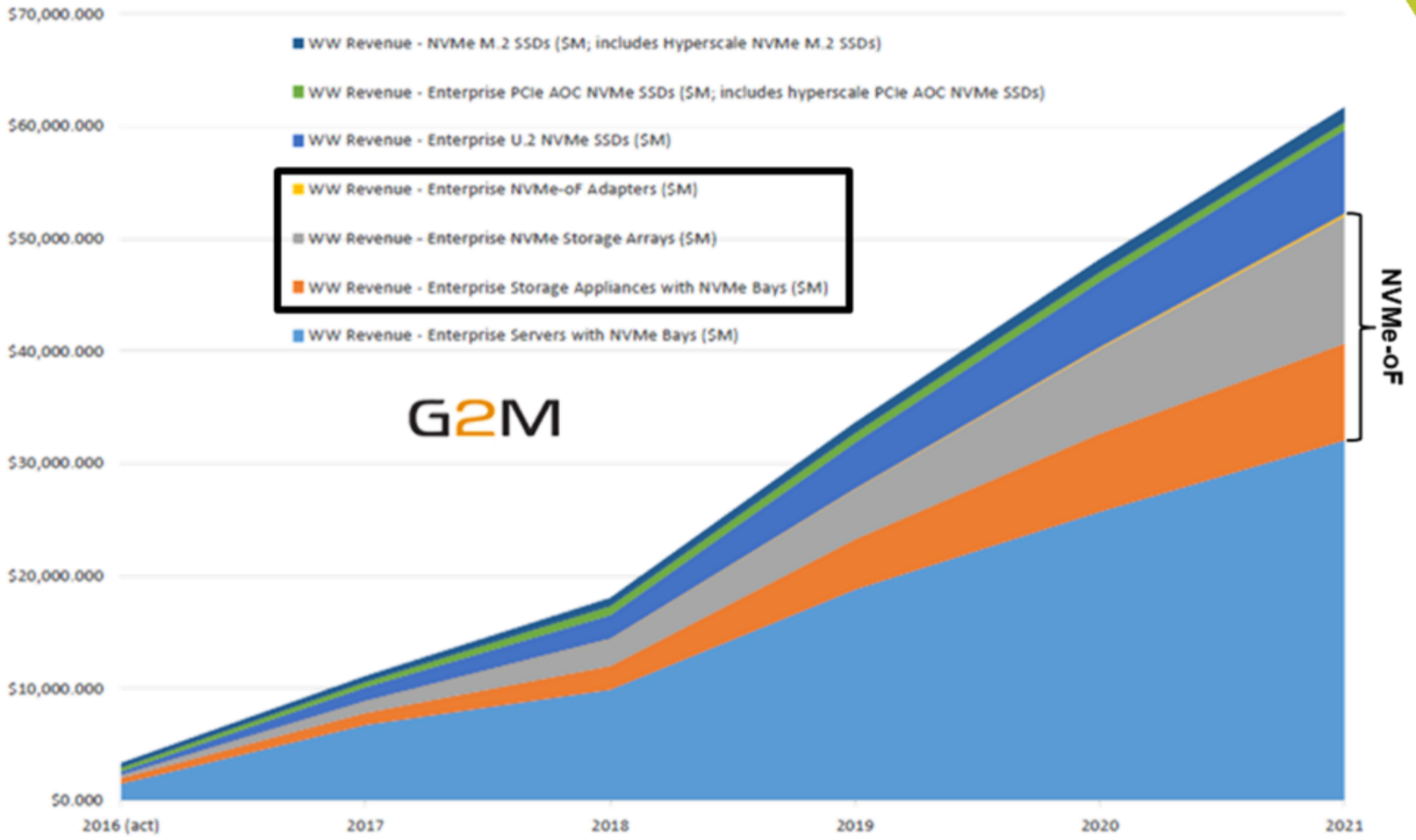


NVMe over Fabrics Maturity

- UNH-IOL, a neutral environment for multi-vendor interoperability since 1988
- Four plug fests for NVMe-oF since May 2017
- Tests require participating vendors to mix and match in both Target and Initiator positions
- June 2018 test included Mellanox, Broadcom and Marvel ASIC solutions
- URL to list of vendors who OK public results: https://www.iol.unh.edu/registry/nvmeof

NVMe Market Projection – $60B by 2021

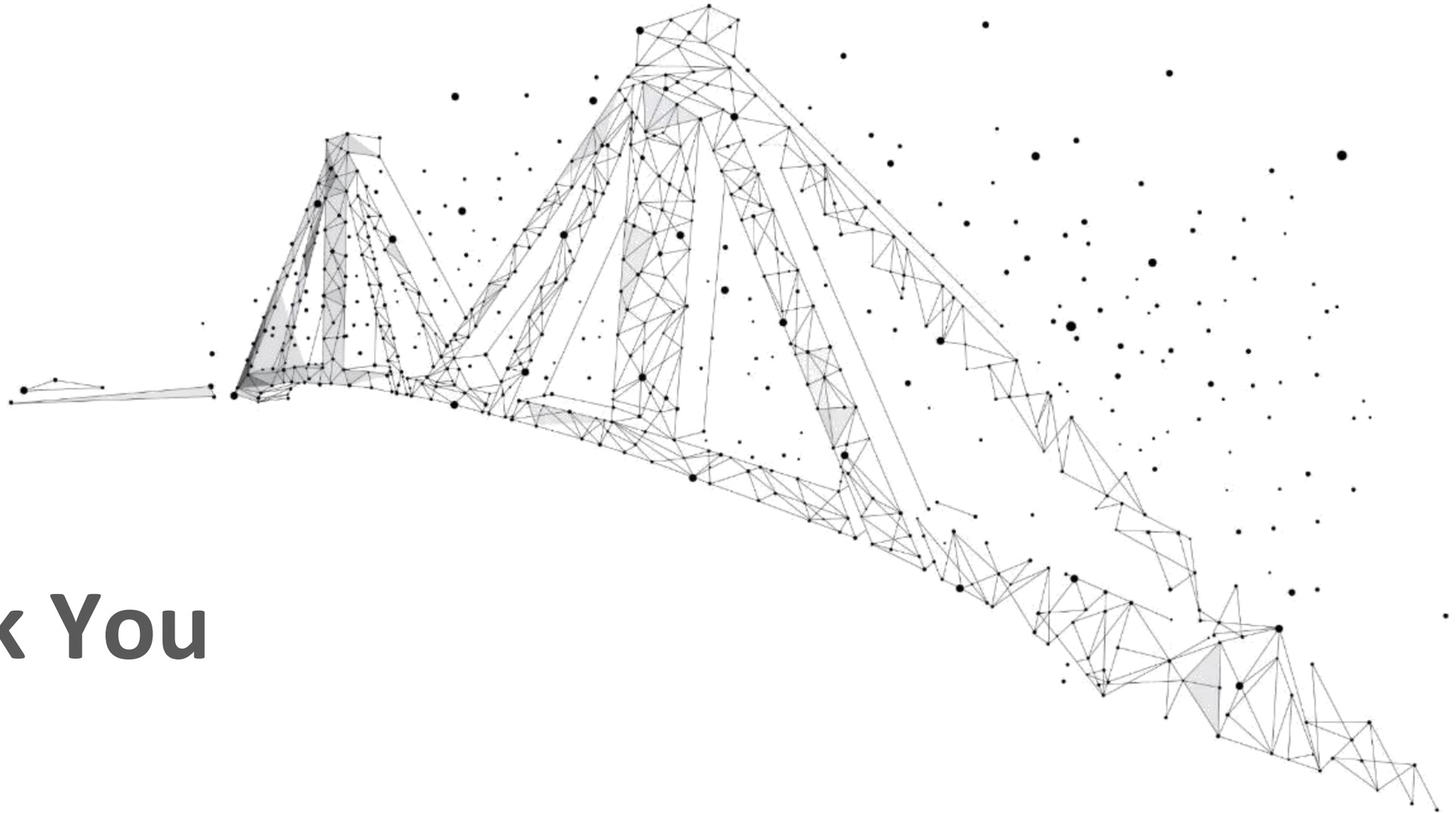
- ~$20B in NVMe-oF revenue projected by 2021
- NVMe-oF adapter shipments will exceed 1.5M units by 2021
  - This does not include ASICs, Custom Mezz Cards, etc. inside AFAs and other Storage Appliances



Some NVMe-oF Storage Players

Conclusions

- NVMe-oF brings the value of networked storage to NVMe based solutions
- NVMe-oF is supported across many network technologies
- The performance advantages of NVMe, are not lost with NVMe-oF
  - Especially with RDMA
- There are many suppliers of NVMe-oF solutions across a variety of important data center use cases



Thank You

SDC 19
SNIA INDIA
May 23-24, 2019
Bangalore, India
STORAGE DEVELOPER CONFERENCE

NVMe over Fabrics Demystified

Rob Davis
Mellanox