NVMe over Fabrics Demystified

Rob Davis
Mellanox
Why NVMe over Fabrics?

Storage Media Technology

Access Time in Micro Seconds

- HDD
- SSD
- PM

10,000x improvement
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 1<sup>st</sup>
  - 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

- SATA HDDs
  - 24 SATA HDDs
  - 100 SATA HDDs

- SATA SSDs
  - 2 SATA SSDs
  - 9 SATA SSDs

- NVMe
  - 1 NVMe
  - 2 NVMe
  - 4 NVMe

Bandwidth (MB/s)

10GbE

© 2019 Mellanox Technologies
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

Efficient Data Movement (RDMA):
- Higher Bandwidth
- Lower Latency
- More CPU Power For Applications

Kernel Bypass
Protocol Offload

Application Buffer Network Buffer

Hardware
Adapter based transport

Application

User

Sockets
TCP/IP
Driver
NVMe/RDMA

- Ethernet
  - RoCE
  - iWARP
- InfiniBand
- OmniPath

NVMe-oF over RoCE

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

NVMe Submission Queue

NVMe Completion Queue

RDMA Send Queue

RDMA Receive Queue

Target NVMe Controller Queue

Target RDMA Queues

Network

Host RDMA Queues

Host NVMe Queues
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

Server or Array

PC
HBA

Eth
NIC
HBA

IB

CPU

Network

JBOF

NVMe SSD

NVMe SSD

NVMe SSD

NVMe SSD

PCIe Switch

CPU

Frontend

Backend
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

Software
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**

### No Offload

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

### Offload

- 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

Host Server
OS/Hypervisor
NVMe Standard Driver

NVMe

Physical Local Storage

Local Physical Storage to Hardware Emulated Storage

NVMe Drive Emulation

Host Server
OS/Hypervisor
NVMe Standard Driver

NVMe

Remote Storage

PCIe BUS
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

[Graph showing latency comparison between Local SSD Write, RDMA Write, TCP Write, Local SSD Read, RDMA Read, and TCP Read.]
Latency: NVMe-RDMA vs NVMe-TCP
Latency: NVMe-RDMA vs NVMe-TCP

Random Read (μs)

- Average: 120
- 99%: 167
- 99.9%: 212

Random Write (μs)

- Average: 47
- 99%: 71
- 99.9%: 95

QD = 1
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
NVMe over Fabrics Demystified

Rob Davis
Mellanox