NVMe over Fabrics - High performance SSDs networked for composable infrastructure

Rob Davis, VP Storage Technology, Mellanox
OCP Evolution...

Physical Rack

Server Density

Compute Storage Disaggregation

Lately called “Composable Infrastructure”

Server Disaggregation
Why NVMe over Fabrics?

Storage Media Technology

Access Time in Micro Seconds

HDD  SSD  PM

10,000x improvement
NVMe Technology Background

- Optimized for flash
  - Traditional SCSI designed for disk
  - NVMe bypasses unneeded layers
  - Dramatically reducing latency and increasing bandwidth

![Comparison of MB/s for different storage options](image)
NVMe over Fabrics Enables Storage Networking of NVMe SSDs

- Sharing NVMe-based storage with multiple servers
  - Better utilization: capacity, rack space, and power
  - Better scalability
  - Management
  - Fault isolation

- While maintaining NVMe Performance
NVMe over Fabrics Technology

- Extends NVMe efficiency over a fabric
  - NVMe commands and data structures are transferred end to end
  - Relies on RDMA for performance
  - Bypassing TCP/IP
Faster Storage Needs a Faster Network

SATA HDDs

SATA SSDs

NVMe

Graphs showing bandwidth comparison for SATA HDDs, SATA SSDs, and NVMe.
Faster Network Wires Solves Some of the Problem...

Ethernet & InfiniBand
End-to-End 25, 40, 50, 56, 100Gb
Going to 200 and 400Gb
Faster Protocols Solves More...
Faster Protocols Solves More…
NVMe, NVMe-oF, and RoCE Protocols
Network Based Offload Engines Complete the Picture

- Protocol offloads
  - NVMe over Fabrics
  - RoCE
- Security offloads
  - Encryption
  - Platform isolation
- Storage offloads
  - Data Integrity
  - Compression

![Graph showing ConnectX-5 Ex NVMeoF Read CPU Usage]

- Hardware
  - 10x Improvement with Offload
- Software

75% CPU Overhead
Only 16% CPU Overhead
Importance of Latency

Ethernet Storage Fabric - ESF

Network hops multiply latency
Storage Platform Latency

**ConnectX no Offload**

- 6M IOPs, 512B block size
- 2M IOPs, 4K block side
- 50% CPU utilization
- ~15usec latency (not including SSD)

**BlueField Offload**

- 8M IOPs, 512B block size
- 5M IOPs, 4K block side
- 0.01% CPU utilization
- ~5usec latency (not including SSD)
Composable Infrastructure

- NVMe over Fabrics enables Composable Infrastructure
  - Low latency
  - High bandwidth
  - Nearly local disk performance

- High performance network components are required
  - ESF
  - 1 usec or less latency
  - Protocol offloads

- Further offloads needed for storage features
  - Security
  - Compression
  - Data integrity
Offloads in the Compute Node Adapter

- Right place for some applications
  - Data in flight encryption
  - Applications vs. storage node decision

- In-line processing is efficient
  - Data must flow through network adapter regardless
  - Minimize need for special software flow
  - Lower latency data path

- Natural place for security boundary
  - Isolated execution – separate domain from host
Advanced NICs Accelerate Storage Apps

Storage
- NVMe
- Over Fabrics
- Powered by RoCE

Video
- Video Streaming

Big Data
- Apache Spark

Efficient Data Transport
- RoCE
- InfiniBand
- RDMA

Security

Network
- QSFP/
  - SFP28
- QSFP/
  - SFP28
- PCIe Gen3/4

Application Acceleration Delivers Infrastructure Efficiency

X86 SW Processing

X86 Security Challenges & Vulnerabilities
- No isolation = Vulnerable Infrastructure
  - Application & security domain are identical
- Poor performance
- Poor scalability
- Vulnerable to DDOS attacks
SmartNICs Accelerate Storage Apps & Security & ...

**Storage**
- NVMe Over Fabrics
- Powered by RoCE

**Video**
- Video Streaming

**Big Data**
- Apache Spark

**Efficient Data Transport**
- RoCE
- InfiniBand
- RDMA

**Security**
- X86 Processing
- SOC Security Benefits
  - Security & workload isolation & offload
  - Logically separated security domains
  - Secure boot & firmware update
  - Workload control & visibility
  - Fully programmable

**Network**
- Application Acceleration Delivers Infrastructure Efficiency

- BlueField SOC
- ConnectX
- PCIe Switch
- DRAM
- QSFP/SFP28
- PCIe Gen3/4
- 64-bit ARM Cores
Network Accelerates Workloads

**Video**
- CDN: For Content Delivery Worldwide

**Acceleration**
- ConnectX-5 100G
- ConnectX-4 Lx 25G

**Storage**
- Spectrum 25, 50, 100G Switches

**Security**
- NVMe over Fabrics

**Big Data**
- Apache Spark

**Offload**
- Smart NIC
- BlueField SoC
NVMe over Fabrics Enables Composable Infrastructure and Much More…

- NVMe over Fabrics
  - Local storage performance across a network

- High performance network components - ESF
  - RDMA
  - Low latency
  - High Bandwidth

- Hardware offload accelerators
  - Protocols
  - Security
  - Storage features

- Composable Infrastructure, Security, Video, Big Data…
Thank You

robd@mellanox.com