Use Cases for NVMe-oF for Deep Learning and HCI Pooling

Nishant Lodha
Director of Technologies, Marvell
Emerging Use Cases for NVMe-oF

- Background and Motivation
- Use Cases by Fabric
- Accelerating Deep Learning
- Scaling Hyperconverged Infrastructure
- Key Takeaways
NVMe-oF Background
NVMe over Fabrics (NVMe-oF)

Industry Standard to scale out NVMe

NVMe-oF is unlocking the value of data

Requires an efficient fabric
Scaling our NVMe Requires a (Real) Network

- Many options, plenty of confusion
- Fibre Channel is the transport for the vast majority of today’s all flash arrays
  - FC-NVMe Standardized in Mid-2017
- RoCEv2, iWARP and InfiniBand are RDMA-based but not compatible with each other
  - NVMe-oF RDMA Standardized in 2016
- NVMe/TCP – is here! Standardized in NOV2018
RDMA Use Cases by Application

- HyperConverged Infrastructure
- VM Migration
- RDMA Accelerated FS
- GPU Direct
- Disaggregated Storage – NVMe-oF
- Disaggregated Storage – iSER
- Disaggregated Storage – SMB
- Low Latency VMs
- FastLinQ
- iWARP
- RoCE
- RoCEv2
- Low Latency VMs
NVMe-oF™ RoCE – Limited Use Cases

Small Scale, Contained and Well Managed Use Cases

- Not Automatic
- Not Precise
- Not for everyone

Keeping the network ‘lossless’
- RDMA/OEFD expertise

RNIC Upgrade Required
- Creates Islands

Congestion
- Skillset Requirements
- Backward Compatibility

Not Automatic
- Not Precise
- Not for everyone

2020 Storage Developer Conference. © Marvell. All Rights Reserved.
Use Cases by Fabric

No one size fits all!

<table>
<thead>
<tr>
<th>DAS, HPC, AI/ML</th>
<th>Enterprise Applications</th>
<th>All Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVMe/RDMA (Ethernet)</td>
<td>FC-NVMe (Fibre Channel)</td>
<td>NVMe/TCP (Ethernet)</td>
</tr>
</tbody>
</table>

- Performance at the cost of complexity
- Leverage existing infrastructure. Reliability is key
- Simplicity is key. Balance of performance and cost

Logos are indicative of workload characteristics only.

2020 Storage Developer Conference. © Marvell. All Rights Reserved.
NVMe-oF: NVMe/TCP

**NVMe over TCP**
- What: Defines a TCP Transport Binding layer for NVMe-oF
- Promoted by Facebook, Google, Intel, Marvell etc.
- Not RDMA-based, Standardized on 2018
- Why:
  - Enables adoption of NVMe-oF into existing datacenter IP network environments that are not RDMA-enabled

**Accelerated NVMe over TCP**
- Addresses Scalability and Congestion challenges with RDMA
- Enables adoption of NVMe-oF into existing datacenter IP network environments

**KEY BENEFITS**
- Ultimate Flexibility - NVMe/TCP and NVMe/RDMA
- Exceptional Performance and full offload NVMe/TCP
- Simplicity of TCP with RDMA like performance
Stack Comparison: Offloaded NVMe/TCP vs. Software NVMe/TCP

**Offloaded NIC**
- NVMeTCP Firmware + TCP Offload

**TCP/IP Stack**
- L2 Driver (qede)

**Any NIC**
- L2 Firmware

**Host**
- NVMe Host (nvme)
- IO stack
  - Send Queue
  - Completion Queue
  - Read/Write Data Buffer

**Host**
- NVMe Host (nvme)
- IO stack
  - Send Queue
  - Completion Queue
  - Read/Write Data Buffer
Reducing Latency, Freeing up CPU

4K Read IO - queued latency [usec]

NVMe/TCP Throughput
Read 8KB, 16Jobs, 8QD

Bandwidth (Gbps)

CPU Utilization (%)

Marvell SW NVMe/TCP
Marvell HW NVMe/TCP

0.00% 20.00% 40.00% 60.00% 80.00% 100.00%

0.00% 20.00% 40.00% 60.00% 80.00% 100.00%

43.4%
16.6%

0.00 5.00 10.00 15.00 20.00 25.00 30.00

2020 Storage Developer Conference. © Marvell. All Rights Reserved.
Use Cases for AI/ML
An AI Breaks the Writing Barrier

A new system called GPT-3 is shocking experts with its ability to use and understand language as well as human beings do

By David A. Price • Aug. 22, 2020 12:01 am ET

WSJ
Deep Learning

- Two step – Training (continuous) and Inference

Where and What

- Compute and Storage intensive
- In Datacenter
  - On or off Prem
- Large Clusters of CPUs, GPUs and NVMe
- Latency sensitive
- In Datacenter
  - Smaller set of CPUs/GPUs or mix
- At the edge
  - CPUs, ASICs, FPGAs
Machine Learning IT Landscape

- Federated Learning
- Machine Learning IT Landscape
- Deep Learning
- Big Data
- Data Analytics
- Edge
- Edge
- Edge
- Edge
- Compute
- Edge
- Device

- Low-latency inference
- Training
- Inference
- Hybrid/Public

- Private
Machine Learning Pipeline

GPU Storage Bottleneck

ML training datasets typically far exceed GPU’s local RAM capacity, creating an I/O chokepoint that analysts call the GPU storage bottleneck. AI and ML systems end up waiting, and waiting, to access storage resources – their massive size impeding timely access and thus performance. AI and ML applications involve accessing a large number of small files from many GPU servers, deploying a parallel distributed file system as the storage infrastructure becomes a necessity. NVMe-oF provides GPUs with direct access to an elastic pool of NVMe, so all resources can be accessed with local flash performance. It enables AI data scientists and HPC researchers to feed far more data to the applications so they can get to better results faster.
GPU to GPU Communications

- 100GbE NIC on NVIDIA DGX-1 Compliant Servers
  - Supports GPUDirect RDMA
    - Lower latency
    - Higher throughput
    - Lower CPU utilization
  - Integrated with NCCL 2.0 (Inter-node Communication over RDMA)
  - Supports distributed TensorFlow / Horovod
Scaling out GPUs over the Network

From scattered, underutilized, non-optimized and uncoordinated GPUs deployment to **Unified, Virtualized and Elastic** GPU cluster
Deep Learning Storage Optimization with NVMe-oF

- **Problem:** Captive storage in Deep Learning nodes
- **Solution:** Storage Pool on EBOF vs. Captive per blade/R&T storage
  - A RDMA fabric provides excellent scalability for CNNs
  - Delivers a high-performance data platform for deep learning, with performance on par with locally resident datasets
  - GPU Direct technology enable direct GPU to GPU communication over RoCEv2
  - Customer can reduce SKUs by disaggregating storage
Use Cases for HCI
VMware vSAN

• With ESXi 6.X/7.X running NVMe over Ethernet

• Key Benefit – low CPU utilization, future-proof configuration

• I/O Capabilities require
  ▪ 10GbE or 10/25GbE
  ▪ SR-IOV, VXLAN, N-VDS(E)
  ▪ Storage offload
  ▪ NVMe/RoCE, NVMe/TCP

* Based on internal Marvell tests
Microsoft Storage Spaces Direct

• What is Storage Spaces Direct (S2D)?
  ▪ Software-defined storage / HCI
  ▪ Highly available and scalable
  ▪ Storage for Hyper-V and Private Cloud
  ▪ Industry standard hardware – servers, storage, networking
  ▪ RDMA (RoCE or iWARP) network as storage fabric
  ▪ 10/25GbE
Scalable HCI

- **Problem:** Storage and Compute tied to the hip
- **Solution:** Shared compute-less storage
  - Marvell NVMe-oF NIC ↔ EBOF’s
  - Utilization: compute nodes are dedicated to run VMs/applications (lower overhead to manage in-node storage [DAS])
  - Scalability: greater networking and storage efficiency with reduced intra-node traffic in the cluster
  - Daisy chaining of EBOFs for scale up deployments
  - Next generation HCI fabrics being enabled to consume external EBOF
  - iWARP for easy deployment and lower OpEx

Improving Storage utilization and scalability For HCI
Key Takeaways
### Many Applications, Many Choices

**No one size fits all!**

<table>
<thead>
<tr>
<th>DAS, HPC, AI/ML</th>
<th>Enterprise Applications</th>
<th>All Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVMe/RDMA (Ethernet)</td>
<td>FC-NVMe (Fibre Channel)</td>
<td>NVMe/TCP (Ethernet)</td>
</tr>
</tbody>
</table>

- **DAS, HPC, AI/ML**
  - NVMe/RDMA (Ethernet)
  - Symbiosis of performance and cost
  - Logos are indicative of workload characteristics only.

- **Enterprise Applications**
  - Leverage existing infrastructure. Reliability is key.
  - Logos are indicative of workload characteristics only.

- **All Applications**
  - Simplicity is key.
  - Balance of performance and cost.

---

Logos are indicative of workload characteristics only.

2020 Storage Developer Conference. © Marvell. All Rights Reserved.
Please take a moment to rate this session.

Your feedback matters to us.