Demonstrating NVMe Over Fabrics Performance on Ethernet Fabrics

Rob Davis
VP of Storage Technology
Mellanox
Why NVMe and NVMe over Fabrics (NVMf)

Storage Media Technology

10,000x improvement

Access Time in Micro Seconds

- HDD
- SSD
- PM
What Makes NVMe Faster

~30μsec

~80μsec
NVMe Performance

- NVMe flash outperforms SAS/SATA flash
  - 2x-4x more bandwidth, 50-60% lower latency, Up to 5x more IOPS
- NVMe is optimized for flash and next-gen persistent memory (PM)
  - Traditional SCSI interfaces designed for spinning disk
  - NVMe bypasses unneeded layers
NVMf is the Logical and Historical next step

- Sharing NVMe based storage across multiple servers/CPUs
  - Better utilization: capacity, rack space, power
  - Scalability, management, fault isolation
- NVMf Standard 1.0 was completed in early June
How Does NVMf Maintain Performance

- The idea is to extend the efficiency of the local NVMe interface over a fabric
  - Ethernet or IB
  - NVMe commands and data structures are transferred end to end
- Relies on RDMA for performance
  - Bypassing TCP/IP
Why not Traditional TCP/IP Network Stack
What is RDMA

Efficient Data Movement (RDMA)

Application
Buffer → Network → Buffer
Kernel Bypass Protocol Offload

User S/W
User Application
IO Library
Kernel S/W
Kernel Apps
OS Stack
Sys Driver
H/W
PCle
Transport & Network (L4/L3)
Ethernet (L1/L0)

Microsoft Storage Spaces Throughput

Throughput (Gbytes/sec)

Without ROCE
Without ROCE
100GbE
ROCE

Standard NIC Flow
RDMA NIC Flow
RDMA & NVMe: A Perfect Fit

1) NVMe command encapsulated into RDMA by host RNIC, then crosses the network.
2) Target moves command to NVMe SSD.
3) Target SSD response encapsulated into RDMA by Target RNIC and crosses the network.
4) Host de-encapsulates response back to NVMe.
Early Pre-standard Demonstrations

- April 2015
  - NAB Las Vegas

- 10Gb/s Reads, 8Gb/s Writes
- 2.5M Random Read 4 KB IOPs
- Latency ~8usec over local
Compute/Storage Disaggregation
Micron FMS 2015 Demonstration

100GbE RoCE

Target Server

Target | Local 4KB read IOPS | Local 4KB write IOPS | Remote read IOPS | Remote write IOPS | Remote write added latency | Remote read added latency
--- | --- | --- | --- | --- | --- | ---
1 NVMe SSD | 849K | 330K | 845K | 330K | 1.9us | 4.76us
4 NVMe SSDs | 3406K | 1333K | 3388K | 1332K | N/A | N/A
MemBlaze Demonstrations – 40GbE/RoCE

![Graph showing PBlaze SSD NBDx performance comparison between 1 SSD and 2 SSDs. The graph indicates a linear increase in IOPS(K) and THROUGHPUT(GB/s) as the number of SSDs increases.](image)
# Pre-standard Drivers Converge to V1.0

<table>
<thead>
<tr>
<th>Demo</th>
<th>NVMe Hardware</th>
<th>Software / Drivers</th>
<th>Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mangstor</td>
<td>Mangstor</td>
<td>Mangstor NVMeoF</td>
<td>RoCE or IB</td>
</tr>
<tr>
<td>PMC Sierra</td>
<td>PMC</td>
<td>PMC NVMeoF</td>
<td>40Gb RoCE</td>
</tr>
<tr>
<td>HGST</td>
<td>HGST</td>
<td>HGST NVMeoF</td>
<td>56Gb InfiniBand</td>
</tr>
<tr>
<td>Micron</td>
<td>Micron</td>
<td>Mellanolx NBDx</td>
<td><strong>100Gb RoCE</strong></td>
</tr>
<tr>
<td>Memblaze</td>
<td>Memblaze</td>
<td>Mellanolx NBDx</td>
<td>40Gb RoCE</td>
</tr>
<tr>
<td>Samsung at FMS15</td>
<td>Samsung</td>
<td>iSER / Ceph / SMB Direct</td>
<td>40Gb RoCE</td>
</tr>
<tr>
<td>Intel at IDF14</td>
<td>Intel</td>
<td>Intel/Chelsio NVMeoF</td>
<td>40Gb iWARP</td>
</tr>
<tr>
<td>Stealth startups</td>
<td>Any / Intel NVMe</td>
<td>Startup’s NVMeoF</td>
<td>40Gb RoCE</td>
</tr>
</tbody>
</table>

![Diagram](image-url)
NVMf Standard 1.0 Community Open Source Driver Development

Mellanox
Intel
HGST
EMC
Apeiron Data Systems
Broadcom Corporation
Chelsio Communications, Inc
Excelero
Hewlett Packard Enterprise
Kazan Networks
Kenneth Okin Consulting
Mangstor
NetApp
Oracle America Inc.
PMC
Qlogic Corporation
Samsung
SK hynix Inc.
Early Community Driver Performance

- **Topology** –
  - Two compute nodes
    - ConnectX4 25GbE RoCE
  - One storage node
    - ConnectX4-LX 50GbE RoCE
    - 4 X Intel NVMe device (P3700/750 series)
  - Nodes connected through switch
- BS = 4k, 16 jobs, IO depth = 64
- 4 cores @ 50% utilization

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>IOPS</th>
<th>Added latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2GB/sec</td>
<td>1.3M</td>
<td>~12us</td>
</tr>
</tbody>
</table>
Kernel & User Based NVMf

- Throughput of NVMf with polling user driver can reach ~1.0M IOPS, with only 1 CPU cores utilized.
Intel NVMf IDF Shenzhen Demo

NVMf Initiator

PCIe 3.0 x8

RNIC

NVMf Target

PCIe 3.0 x8

RNIC

NVMe SSDs

40GbE with RDMA (RoCE)

IOPS (million I/Os per sec)
Number of cores utilized
% CPU Utilization (36 core/72 HT Xeon platform)
Conclusions

- Future Storage solutions will be able to deliver DAS storage performance over a network if:
  - NVMe SSDs – new NVMe protocol eliminates HDD legacy bottlenecks
  - Fast network – “Faster storage needs faster networks!”
  - NVMf with RDMA – new NVMf protocol running over RDMA is within microseconds of DAS
Thanks!

Rob Davis
robd@mellanox.com