NVMe over Fabrics
Storage’s New Magic Wand
Arindam Sarkar
MSys Technologies LLC
About MSys Technologies

Our WW Strength
800
And growing

Product Engineering Services
- Storage and Networking Engineering
- VMware Ecosystem Integration
- UX/UI, Enterprise Mobility
- Rapid Application Development
- AI, ML and Cognitive Services
- Fintech and Loyalty
- Contingent Hiring

Technology CoEs
- Storage CoE
- DevOps CoE
- QA Automation CoE
- Big data and Predictive Analytics CoE
- Digital Testing CoE
- Cloud CoE
- Open Source CoE

Outsourcing Partners to

Key Alliances

- Hewlett Packard Enterprise
- CHEF
- Rubrik
- Google
- IBM

- Cloud Native Computing Foundation
- Amazon Web Services
- DOCKER
- CHEF
- IBM Watson
Trends in NVMe over Fabrics (1/2)

SDS enables end-to-end NVMe-OF supporting any storage

- 60% of Software Defined Storage servers will have NVMe bays by 2020
- SDS server will register growth due to the support of RDMA's for OpenStack and other SDS platforms
- OS and Hypervisor vendors are leading the charge to native SDS solutions
- The external arrays will be challenged by NVMe-oF and Hyper-Convergence
• NVMe-OF JBOFs are replacing DAS
• NVMe-OF are enabling vendors to define new architectures
• Adoption of AFAs and NVMe Storage driving the need for faster networks
• Rack-scale shared storage solution scales to hundreds of NVMe devices
Drivers of Adoption – NVMe/NVMe-OF

**High Performance Computing**
- Database & OTLP
- Oracle, NoSQL, Mem SQL

**Telco NFV**
- IMDB & Analytics
- HANA and Hekaton

**IoT Fog Computing**
- Scale Out SD Storage & Fibre Channel Lives

**Vertical**

**Enterprise**
- Big Data & Advertising
- Content Distribution & Media Services

**Cloud/xSP**
- Deep Learning & AI Systems

**Flash Memory Summit**
Real-Time Apps demand faster fabrics

Real – Time Applications: The Next Phase of Digital Transformation

Artificial Intelligence  Machine Learning  Real-Time Analytics

All demand lower latency and higher performance from faster fabrics and faster media
• SDS provide increased performance and utilization, reduced down-time cost and management complexity
• NVMe-OF meets the above demand of SDS solutions by sharing NVMe based storage across multiple servers
• SDS enables Cloud Native Data Services for MongoDB, Cassandra, and HDFS
• SDS enables Replication, RAID, Fault-domain aware placement, snapshot, placement control for performance
Benefits:
- Storage services (dedup, compression, thin provisioning)
- High availability at the array
- Fully supported from the array vendor
- Example: NetApp/IBM

Benefits:
- High performance storage
- Lower cost that storage arrays, minimal storage services
- Roll-your-own support model
- Ex. SUSE on Servers configured to be storage targets

Benefits:
- Very low latency
- Low cost
- Great for a single rack/single switch
- Leverages NICs, smart NICs, and HBAs for NVMe-oF to PCIe/NVMe translation
Next Generation of Cloud Storage

- Cloud is embracing the use of networked NVMe capacity
- NVMe-OF for cloud workloads (AI & Analytics) increase scalability and elasticity
- Disaggregation of high performance NVMe storage allows performance and features to scale independently
- NVMe over Fabrics enable NVMe SSDs to scale from a few SSDs to thousands of NVMe SSDs
- Microsoft Azure data centers leverage NVMe SSDs consistent performance of SATA based SSDs
• Rapid deployments of multi-core servers densely packed with VMs and increased adoption of all-flash storage arrays driving the need for high performance storage networking
• Ethernet options –ROCE and iWARP with custom drivers on host side
• Most CSPs implement Ethernet networking for storage
• Scale-out Storage and HCI increasingly adopting Ethernet Storage Fabric
• Data Centers adopting Lossless Ethernet switches with Data Center Bridging (DCB)
• RoCE – 2010 Ethernet specifications improve performance of on-prem & cloud deployments

Ethernet NVMe-oF
- Ethernet with RDMA will be over 70% of shipments
- Scale-out SDS will use NVMe to challenge arrays
- Mellanox is leading with RDMA/RoCE. iWARP is TCP/IP based RDMA
- Broadcom, Chelsio have announced products

Fibre Channel NVMe-oF
- Life extension for Fibre Channel & legacy Storage
- Broadcom, Brocade and Cavium look to 2017 GA
- NVMe-OF uses FCP for data (does back-to-back DMA)

InfiniBand NVMe-oF
- Mellanox ConnectX cards support NVMe-OF using RoCE or TCP
- Given their storage cluster inter-connect business this could be interesting
- IB provides native RDMA
NVMe-OF based Solutions

OpenStack and NVMe over fabrics

MongoDB on Pure Storage Flash Array
NVMe-OF based Solutions (contd.)

Excelero – NVMeOF with HCI

WDC OpenFlex Storage Architecture

Hyper-Converged Infrastructure

- Compute
- Network
- Flash
- HDD

Software Composable Infrastructure

- Compute
- Compute
- Network
- Network
- Flash
- Flash
- HDD
- HDD

VS.

Disaggregated Resources

OpenFlex delivers 40% lower TCO* through precise, on-demand resource allocation
NVMe-OF Performance with Open Source Linux Drivers

OpenFabrics Alliance Workshop 2017
## Benchmarking Test Setup (MSys)

### Setup

<table>
<thead>
<tr>
<th>Hardware:</th>
<th>Software stack:</th>
<th>Tool:</th>
</tr>
</thead>
</table>
| 1.64 core x86_64 host and target systems  
2.64GB RAM  
3.100GB Ethernet ConnectX-4 NICs | 1. Linux NVMe host and target software stack with kernel 4. 10+  
2. 2.250GB null target, 4K queue depth, 64 MQs, single LUN or namespace  
3. NULL block driver with multiple queues for fabric performance characteristics | 1. Fio  
2. 16 jobs, 256 queue depth  
3. 70% write, 30% read |

# fio --bs=32k --numjobs=16 --iodepth=256 --loops=1 --ioengine=libaio --direct=1 --invalidate=1 --fsync_on_close=1 --randrepeat=1 --norandommap --time_based --runtime=60 --filename=/dev/nvme0n1 --name=read-phase --rw=randread

### Benchmarking
1. After establishing a connection between NVMF host (initiator) and NVMF target, find a new NVMe block device in the initiator
2. Perform a simple fio traffic test on the block device for different block sizes
Random R/W (30-70) Latency Tests (MSys)

1. 20 times lower latency compare to iSCSI-TCP upto 4K IO Size
2. 10 times lower latency compare to ISER for 8K and higher
3. 2 times lower latency compare to iSER for all IO size
4. Block layer MQ support come natively to NVMe
1. 20 times lower latency compare to iSCSI-TCP up-to 4K Size
2. 4 times higher IOPs compare to iSER for size 8K and higher.
2020 Predictions (not too farsighted)

- **The NVMe market will be over $57 Billion by 2020**

- **Over 50% of servers will ship with NVMe drives by 2020**

- **Over 60% of storage servers drives are NVMe by 2020**

- **NVMe-oF adapter shipments exceed 740K units by 2020**

- **Over 40% of AFAs arrays will ship NVMe by 2020**

- **NVMe technology will contribute more than 50% revenue to the primary storage market.**

- **40% of All-Flash Arrays will ship NVMe by 2020**

- **30% of NVMe Array Vendors will Q custom flash modules**

- **NVMe Arrays will leverage SDS to provide file system capacities**

- **M.2 Form Factor SSDs will also be used in NVMe based arrays**

- **NVMe Flash Arrays will set the new standard for high performance and low latency**

- **NVMe Arrays may or may not use NVMe-oF adapters if they export files systems via RNICs**

Copyright: G2M Incorporated. All Rights Reserved
Thank You!

Arindam Sarkar

Storage Solutions Architect
Arindam@msystechnologies.com