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

Hewlett Packard Enterprise · CHEF · rubrik · Google · IBM

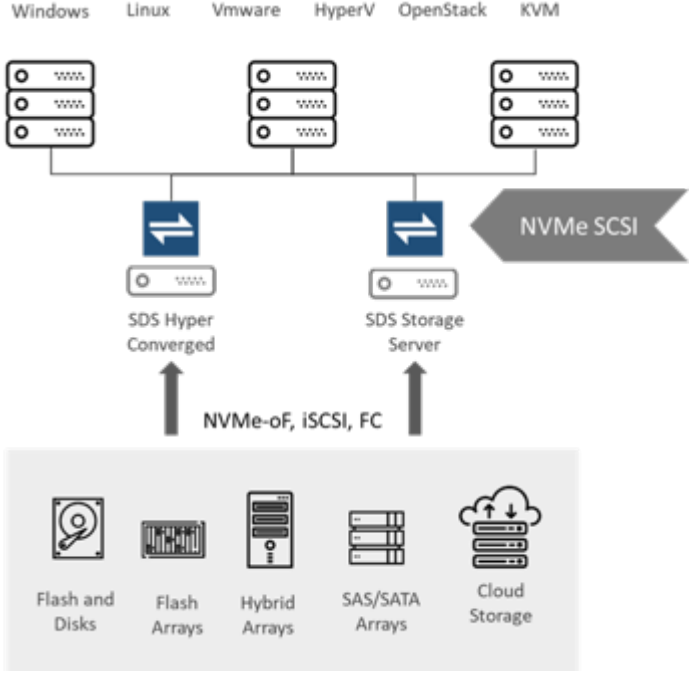
Key Alliances

CLOUD NATIVE COMPUTING FOUNDATION · amazon web services · kubernetes CERTIFIED · docker · CHEF · IBM Watson

Flash Memory Summit

Trends in NVMe over Fabrics (1/2)

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

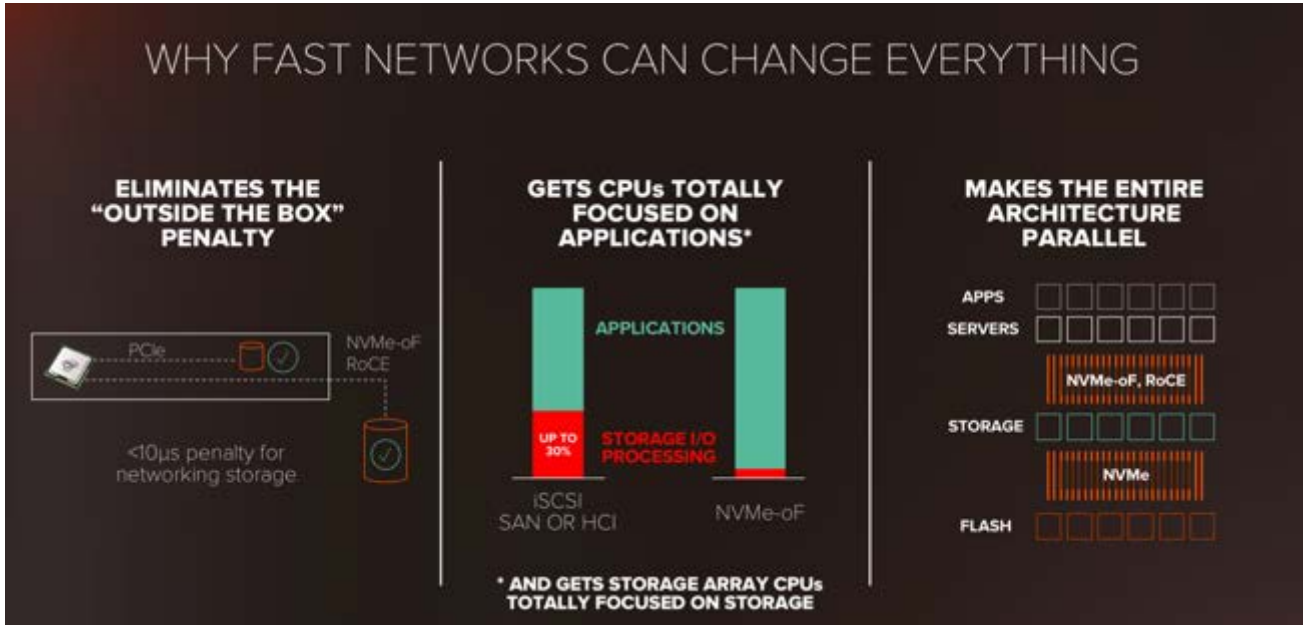


- 60% of Software Defined Storage servers will have NVMe bays by 2020
- SDS server will register growth due to the support of RDMA's for OpenStack and other SDS platforms
- OS and Hypervisor vendors are leading the charge to native SDS solutions
- The external arrays will be challenged by NVMe-oF and Hyper-Convergence

vmware
Microsoft
redhat.
SUSE
openstack
NUTANIX
simplivity

MSys Technologies
Expanding Horizons

Trends in NVMe over Fabrics (2/2)



- 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

Flash Memory Summit

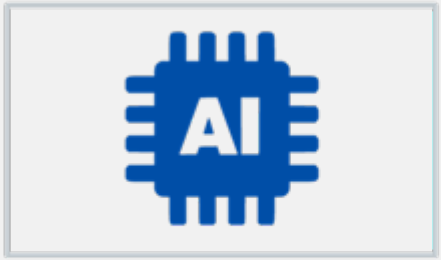
Drivers of Adoption – NVMe/NVMe-OF

	High Performance Computing	Telco NFV	IoT Fog Computing
Vertical			
	DataBase & OTLP Oracle, NoSQL, Mem SQL	IMDB & Analytics HANA and Hekaton	Scale Out SD Storage & Fibre Channel Lives
Enterprise			
	Big Data & Advertising	Content Distribution & Media Services	Deep Learning & AI Systems
Cloud/xSP			

Real-Time Apps demand faster fabrics

Real – Time Applications: The Next Phase of Digital Transformation

Artificial Intelligence



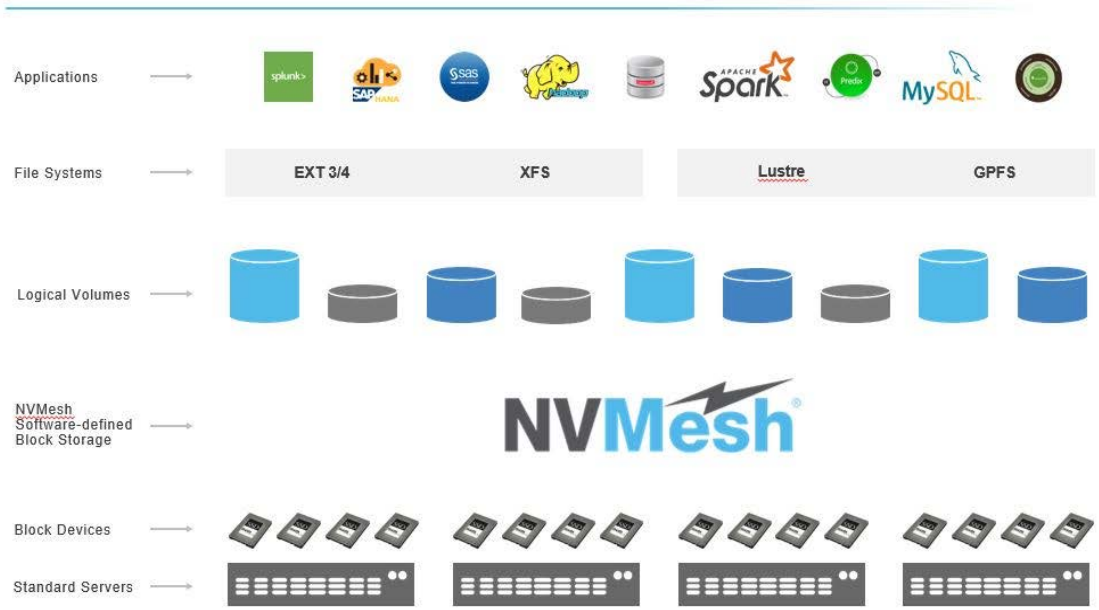
Machine Learning

Real-Time Analytics

All demand lower latency and higher performance
from faster fabrics and faster media

NVMe –Accelerating SDS apps

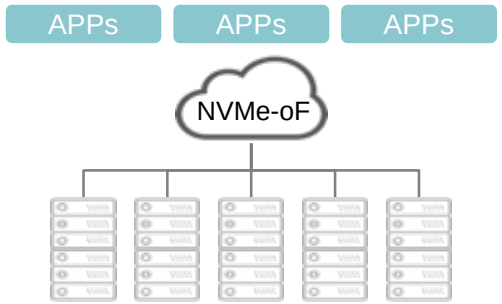
- SDS provide increased performance and utilization, reduced down-time cost and management complexity
- NVMe-OF meets the above demand of SDS solutions by sharing NVMe based storage across multiple servers
- SDS enables Cloud Native Data Services for MongoDB, Cassandra, and HDFS
- SDS enables Replication, RAID, Fault-domain aware placement, snapshot, placement control for performance



NVMesh – Excelero's Softeare Defined Implementation

NVMe-OF –Storage Architectures

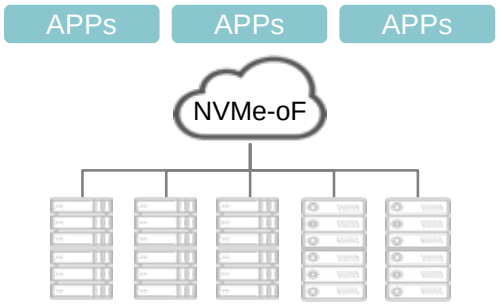
Enterprise Arrays – Traditional SAN



Enterprise Arrays

Benefits:
- Storage services (dedup, compression, thin provisioning)
- High availability at the array
- Fully supported from the array vendor
- Example: NetApp/IBM

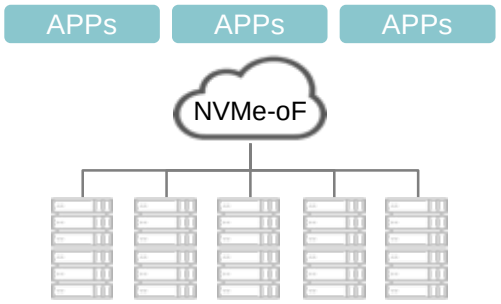
Server SAN/Storage Appliances



Rows of Servers

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

JBOF / Composable Storage

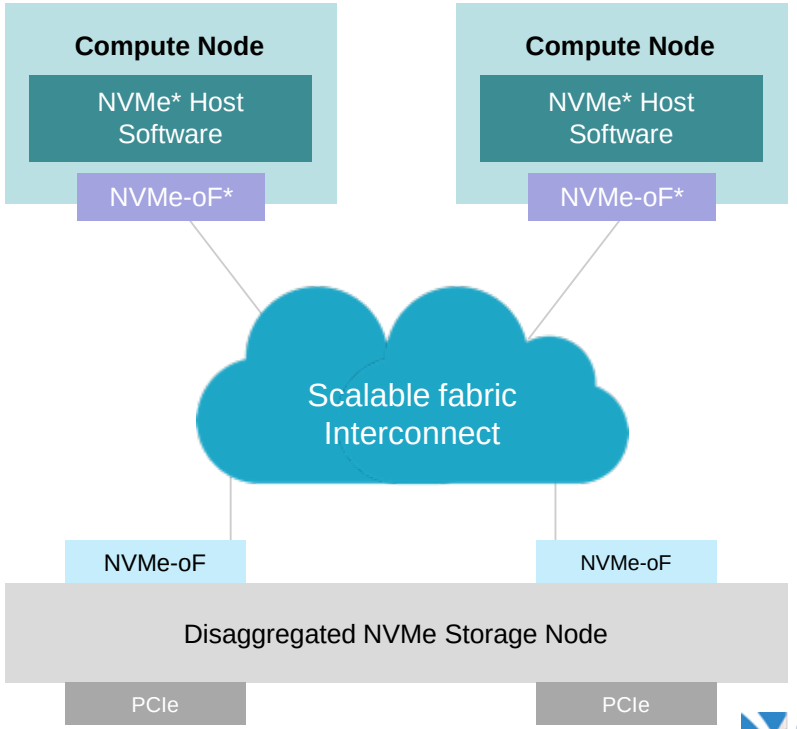


Blocks of Storage

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



NVMe-OF Transports -Ethernet, FC and IB

- 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

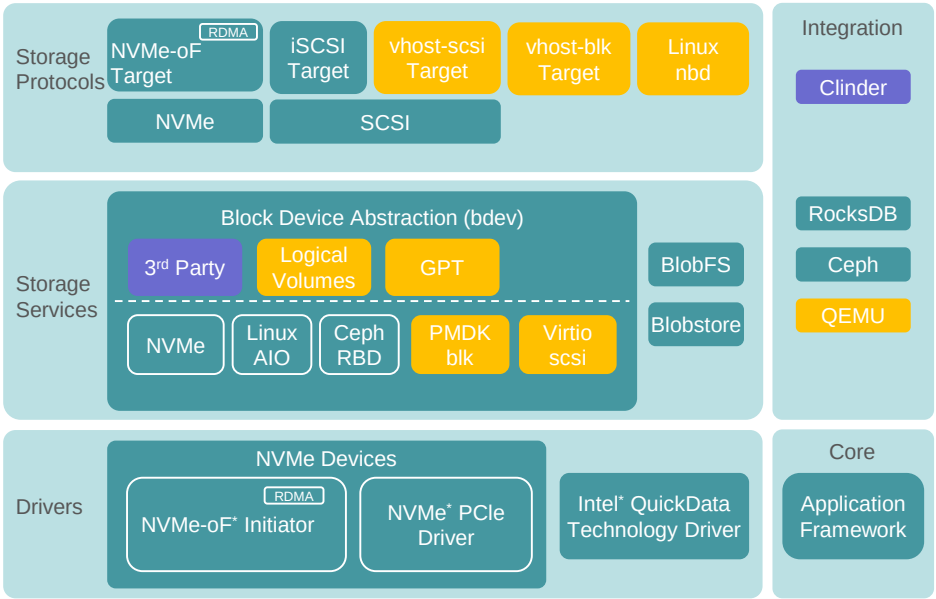
- Life extension for Fibre Channel & legacy Storage
- Broadcom, Brocade and Cavium look to 2017 GA
- NVMe-OF uses FCP for data (does back-to-back DMA)

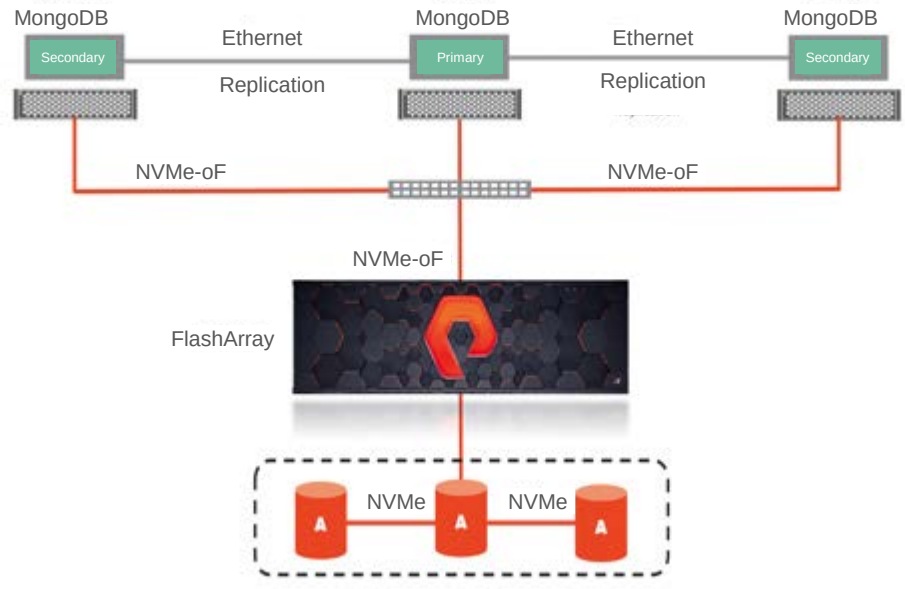
InfiniBand NVMe-oF

- 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

Flash Memory Summit

NVMe-OF based Solutions





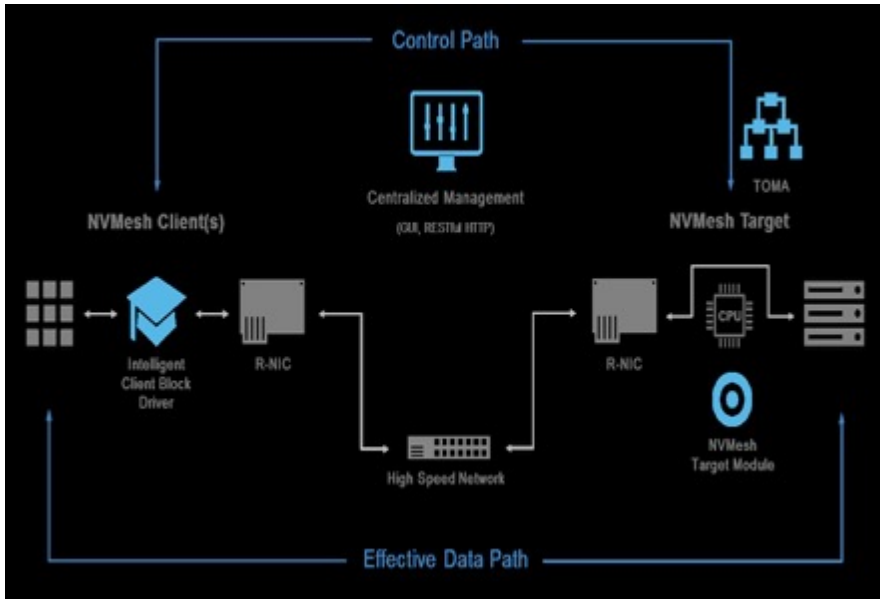
MongoDB on Pure Storage Flash Array

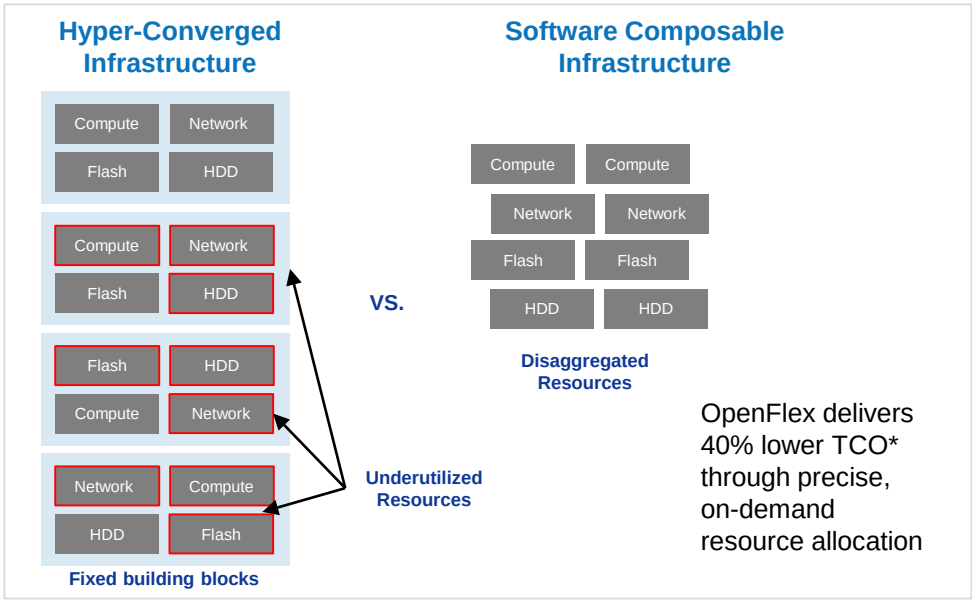
OpenStack and NVMe over fabrics

MSys Technologies
Expanding Horizons

Flash Memory Summit

NVMe-OF based Solutions (contd.)





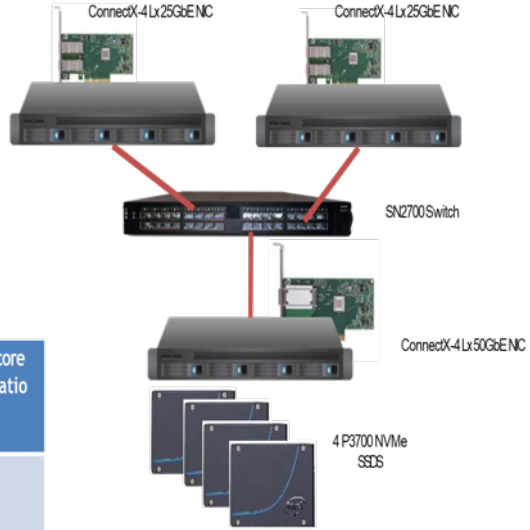
Excelero –NVMeOF with HCI

WDC OpenFlex Storage Architecture

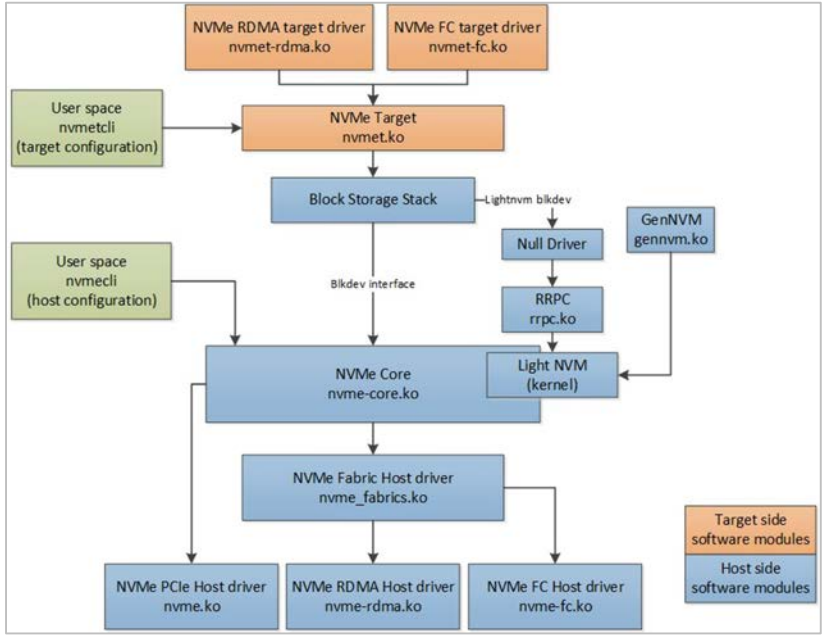
MSys Technologies
Expanding Horizons

NVMe-OF Performance with Open Source Linux Drivers

Added fabric latency
~12us, BS= 512b



	Bandwidth (Target side)	IOPS (Target side)	Num. Online cores	Each core utilization
BS= 4KB, 16 jobs, IO depth = 64	5.2GB/sec	1.3M	4	50%



Flash Memory Summit

Benchmarking Test Setup (MSys)

Setup

Hardware:
1.64 core x86_64 host and target systems
2.64GB RAM
3.100GB Ethernet ConnectX-4 NICs

Software stack:
1.Linux NVMe host and target software stack with kernel 4. 10+
2.250GB null target, 4K queue depth, 64 MQs, single LUN or namespace
3.NULL block driver with multiple queues for fabric performance characteristics

Tool:
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
```

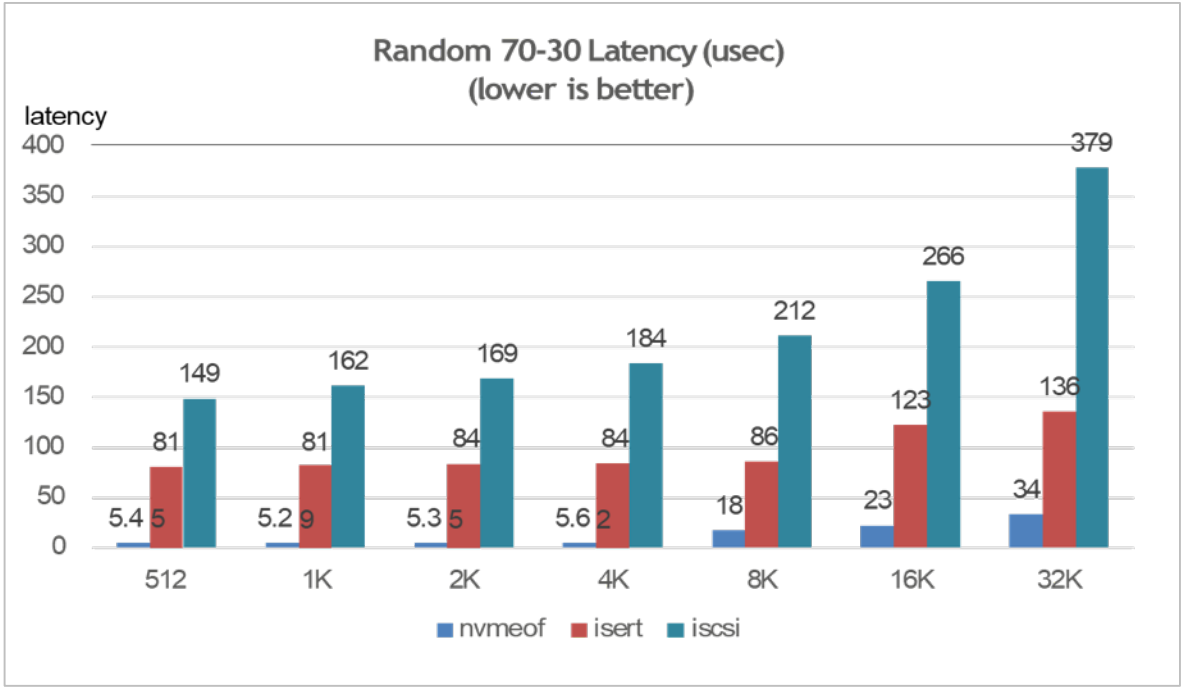
Benchmarking
1. After establishing as connection between NVMF host (initiator) and NVMF target, find a a new NVMe block device in the initiator
2. Perform a simple fio traffic test on the block device for different block sizes

MSys Technologies
Expanding Horizons

Flash Memory Summit

Random R/W (30-70) Latency Tests (MSys)

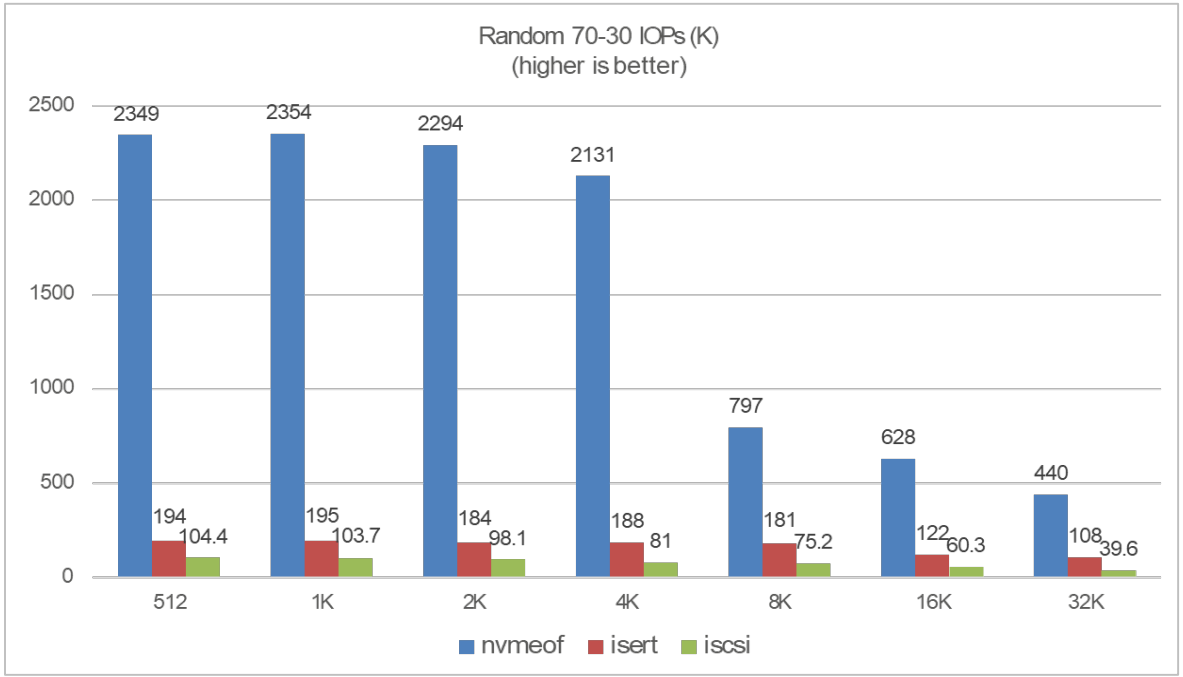
1. 20 times lower latency compare to iSCSI-TCP upto 4K IO Size
2. 10 times lower latency compare to ISER for 8K and higher
3. 2 times lower latency compare to iSER for all IO size
4. Block layer MQ support come natively to NVMe



MSys Technologies
Expanding Horizons

Random R/W (30-70) Latency Tests (MSys)

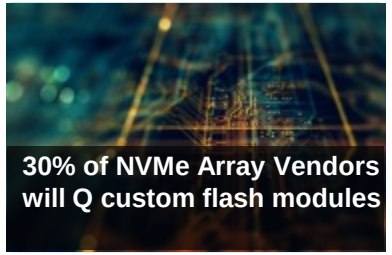
1. 20 times lower latency compare to iSCSI-TCP up-to 4K Size
2. 4 times higher IOPs compare to iSER for size 8K and higher.

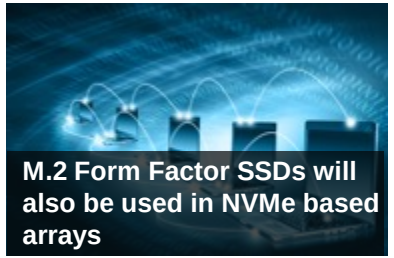


Flash Memory Summit

2020 Predictions (not too farsighted)

NVMe Market Size	The NVMe market will be over $57 Billion by 2020
NMVe SSD U.2 & M.2 in Servers	Over 50% of servers will ship with NVMe drives by 2020
SDS Storage Servers	Over 60% of storage servers drives are NVMe by 2020
NVMe-oF Networking	NVMe-oF adapter shipments exceed 740K units by 2020
AFA Moves to NVMe	Over 40% of AFAs arrays will NVMe based by 2020
NVMe will Dominate	NVMe technology will contribute more than 50% revenue to the primary storage market.





Thank You!

Arindam Sarkar

Storage Solutions Architect
Arindam@msystechnologies.com