

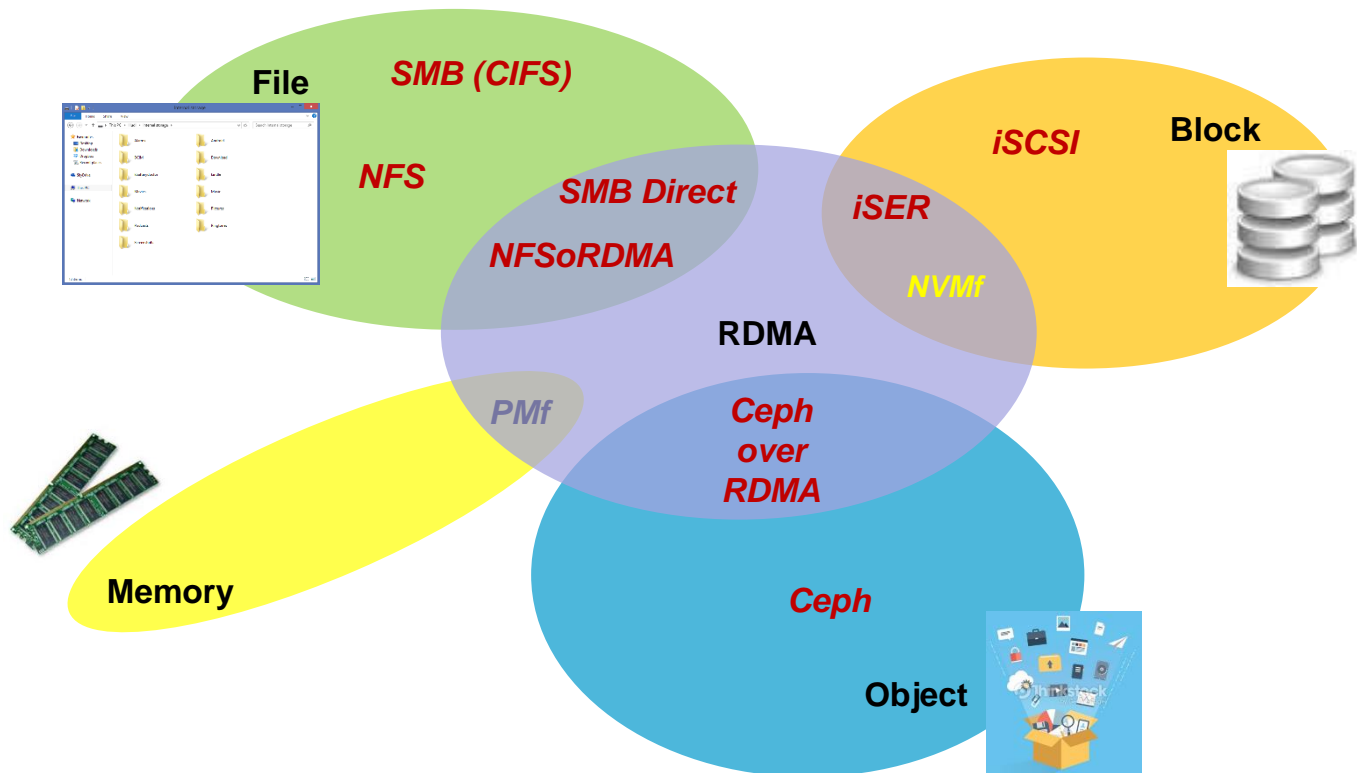
Accelerating Flash Storage with Open Source RDMA

Rob Davis

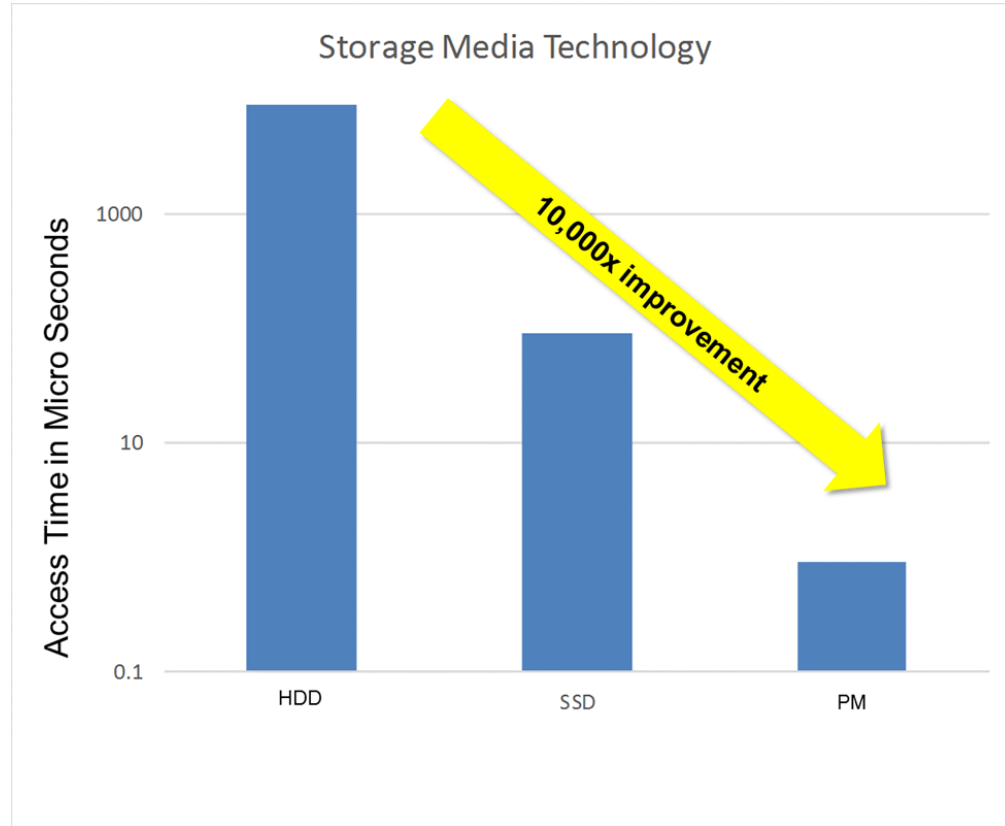
Vice President of Storage Technology

Open Source Flash Storage Solutions

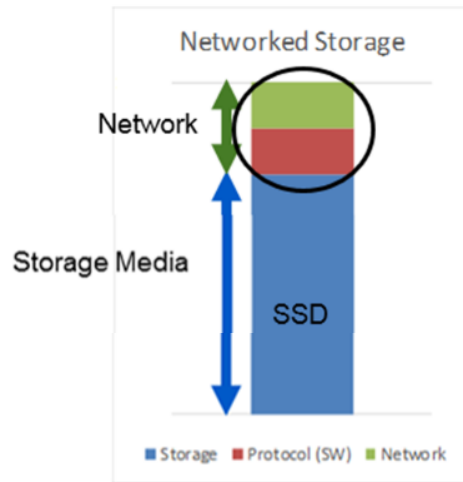
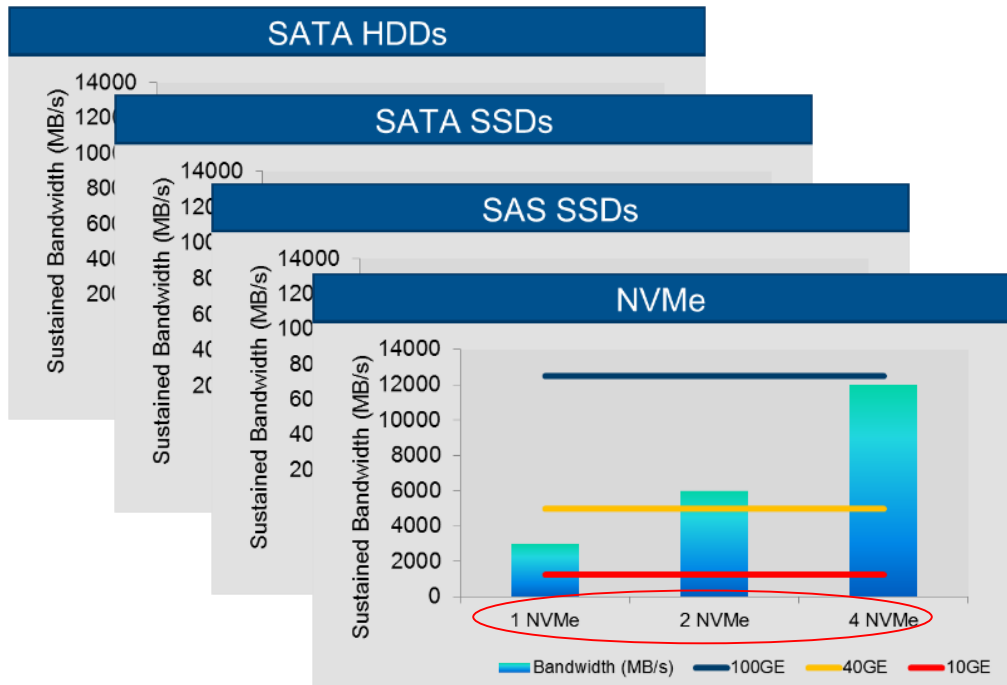
- Pure Bandwidth – up to 100Gb/s
 - Flash over Block, File and Object
- RDMA
 - RoCE, iWARP, InfiniBand
 - iSER
 - SMB Direct, NFSoRDMA
 - Ceph over RDMA
- Non-Volatile Memory (NVM)
 - NVMe over Fabrics (NVMeF)
 - PMf (3D-XPoint)



Why Should We Care About RDMA



Faster Storage Needs a Faster Network



Flash SSDs move the Bottleneck from the Disk to the Network

Faster Wires are Here Today

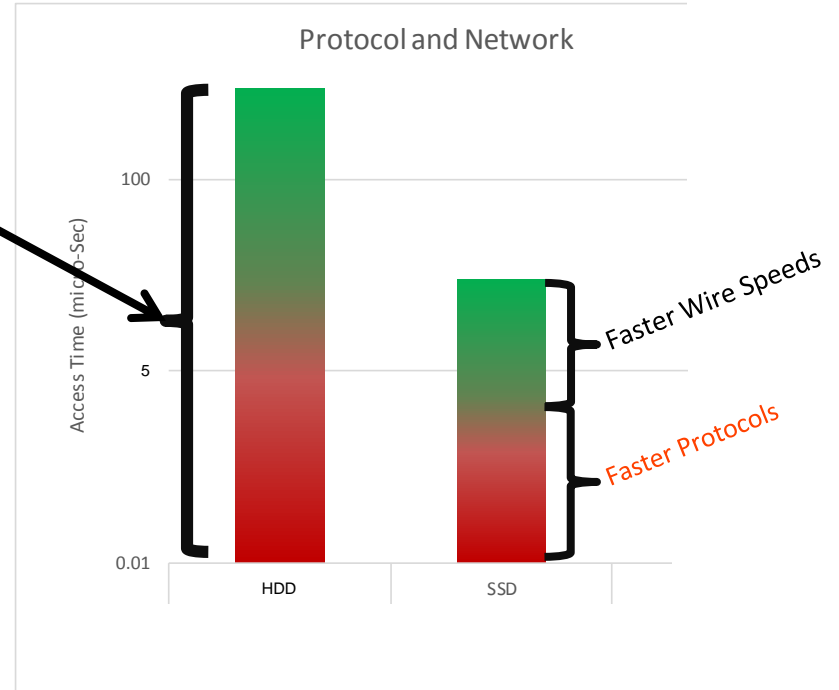
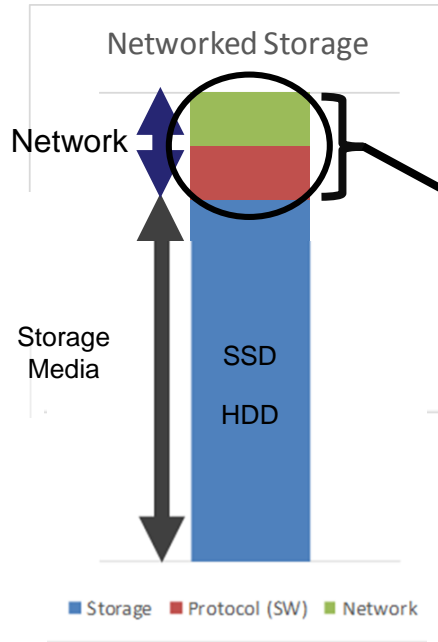


End-to-End 25, 40, 50, 100Gb Ethernet

100Gb InfiniBand

Gen4 PCIe and 32GbFC

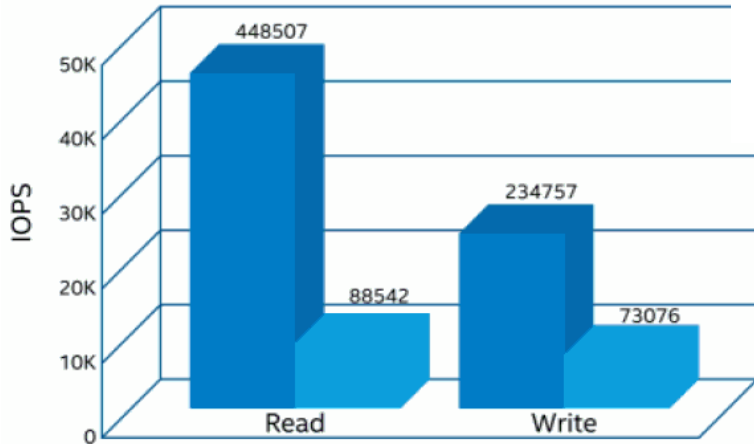
Faster Wires Only Solve ½ the Problem



Faster Protocol: NVMe

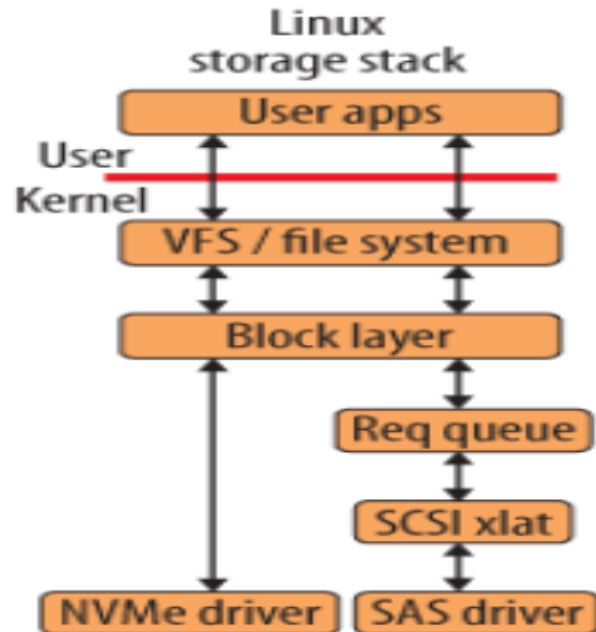
- NVMe: Optimized for flash and next-gen NV-memory
 - Traditional SCSI interfaces designed for spinning disk
 - NVMe bypasses unneeded layers
- NVMe Flash Outperforms SAS/SATA Flash
 - 2x-2.5x more bandwidth, 40-50% lower latency, Up to 3x more IOPS

Random Read/Write Performance[†]
750 Series (PCIe) vs. 730 Series (SATA)



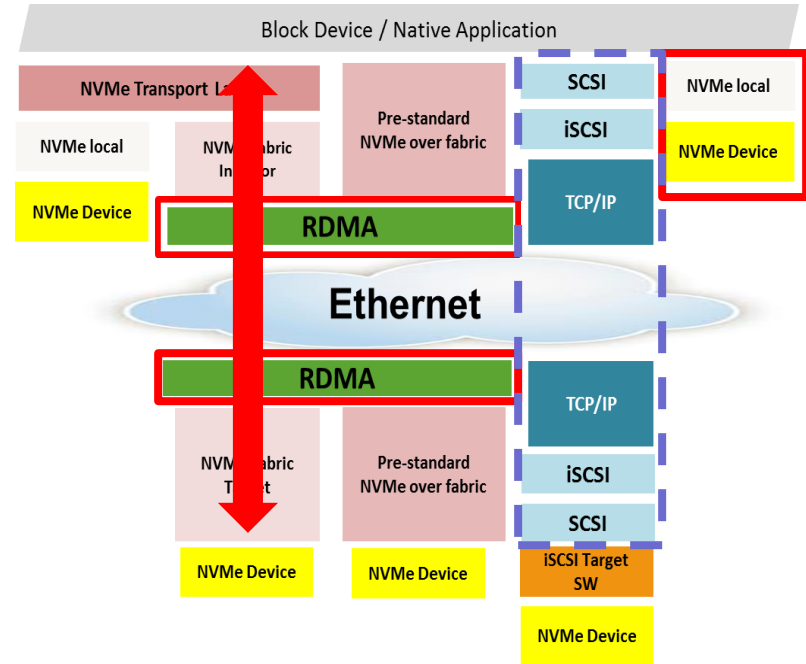
Flash Memory Summit
Santa Clara, CA

■ 750 Series (PCIe) 400GB ■ 730 Series (SATA) 480GB

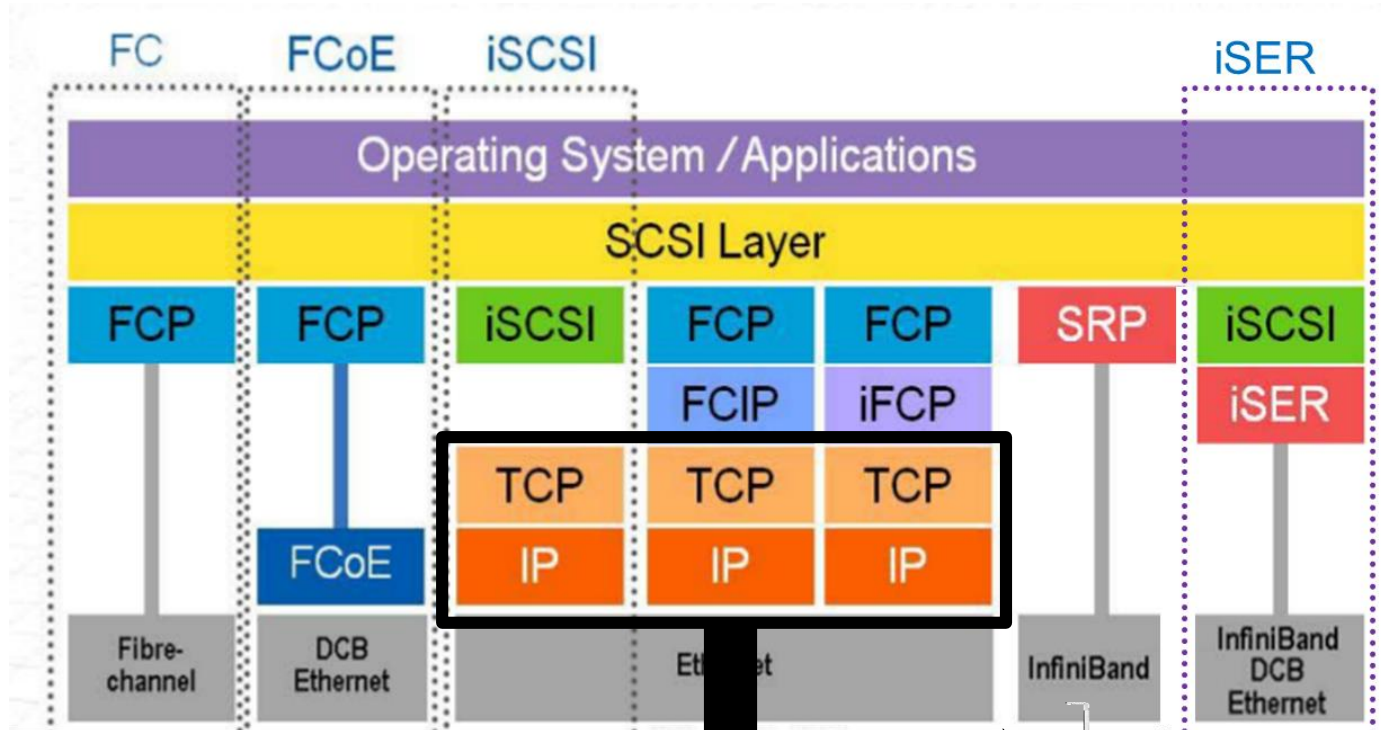


Faster Protocol: NVMeF

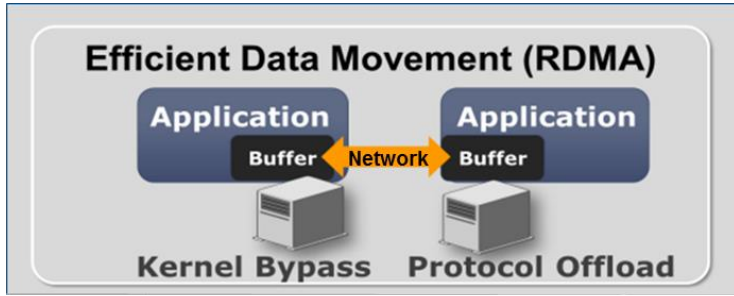
- The idea is to extend the efficiency of the local NVMe interface over a fabric
 - Ethernet or IB
 - NVMe commands and data structures are transferred end to end
- Relies on RDMA for performance
 - Bypassing TCP/IP



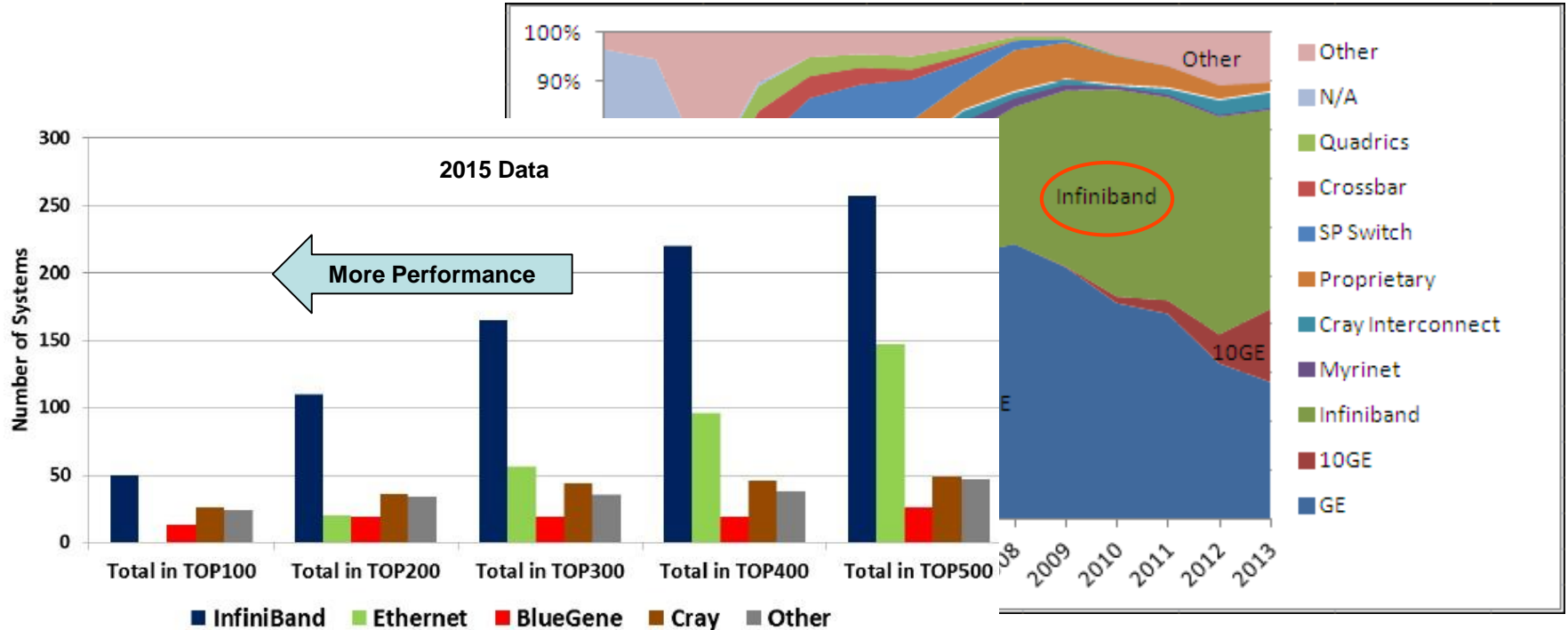
Stack Bypass and Efficiency



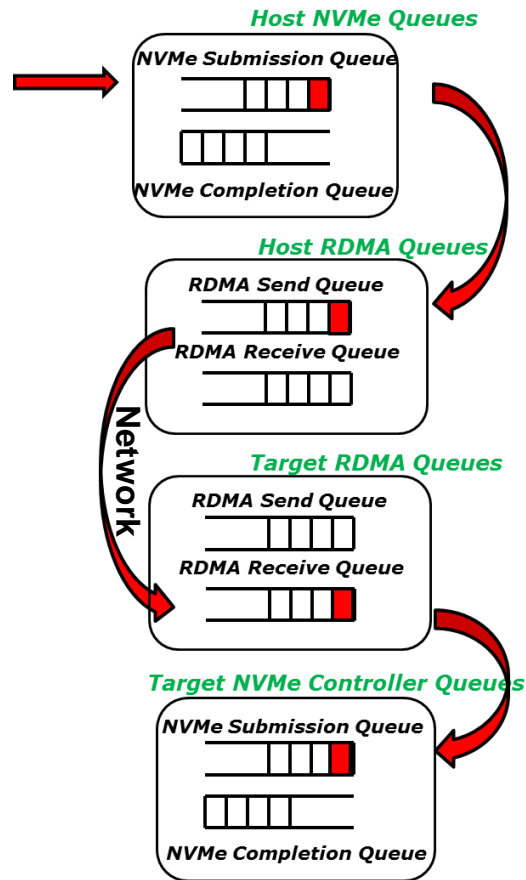
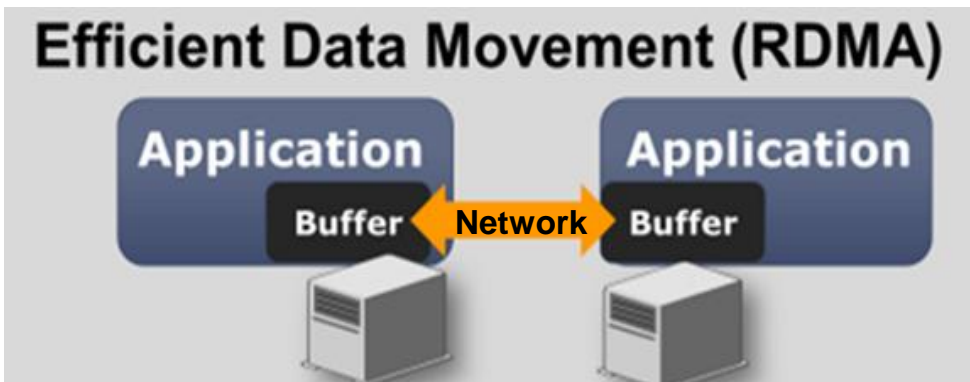
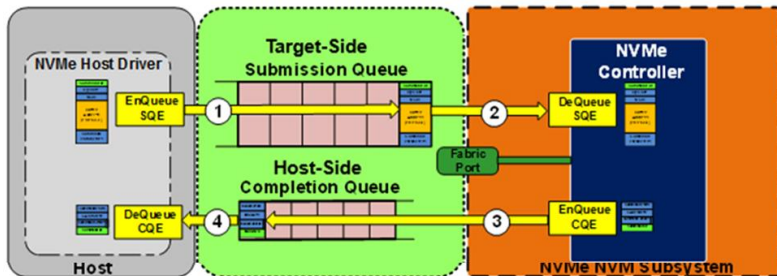
What is RDMA?



RDMA borrowed from HPC

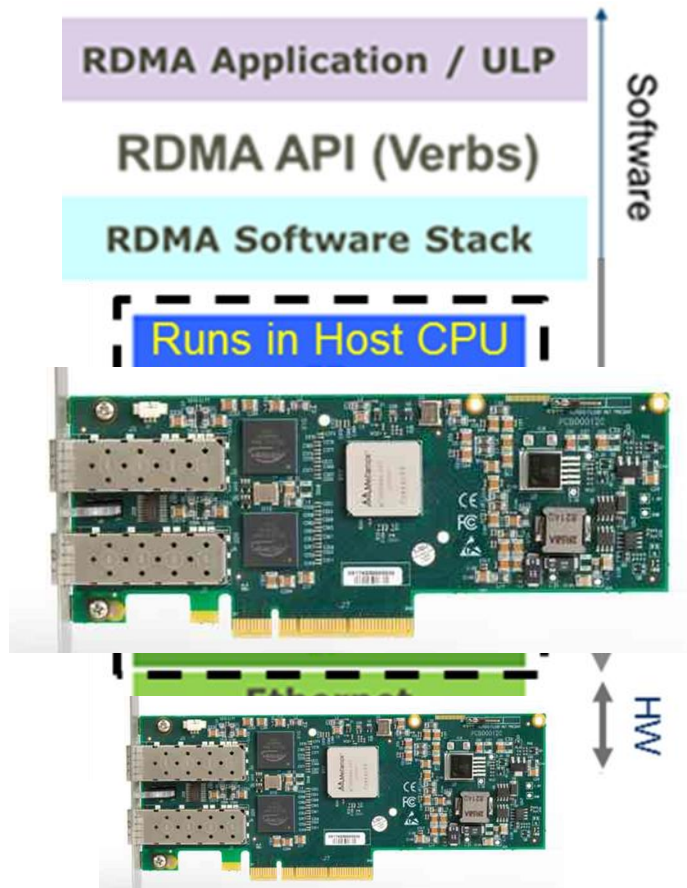


Efficiency



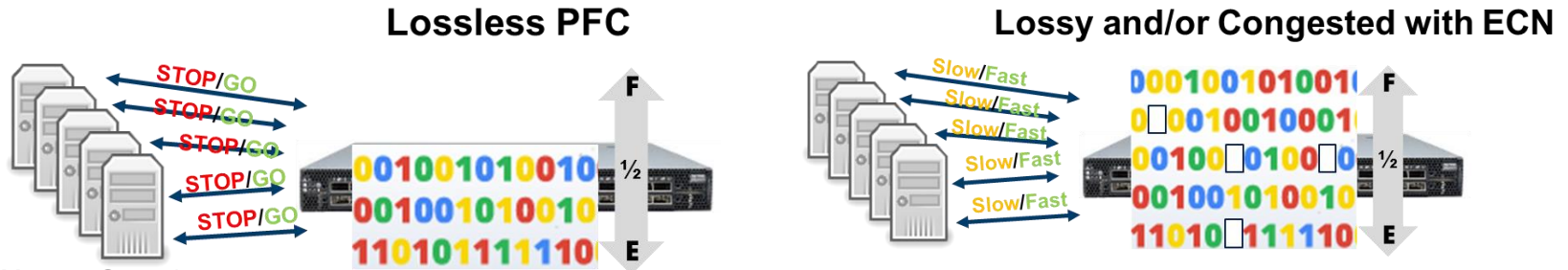
Soft RoCE

- Software implementation of RoCE
- A Driver which can run over any Ethernet NIC
- Fully interoperable with Hardware RoCE devices
- Benefits
 - Heterogeneous RoCE networks
 - Roll out hardware acceleration in stages
 - Accelerate RoCE deployments
 - Easier RoCE testing and development
- Open Source



Resilient RoCE Update

- Not a new standard or version of RoCE – implementation enhancement
- Most of today's RoCE products require a lossless network implemented through PFC(IEEE standard Priority Flow Control)
- Some Data Center prefer ECN over PFC on their networks
 - IEEE standard ECN(Explicit Congestion Notification)
- This Update enables running RoCE on a Lossy and/or Congested network with ECN with or without PFC



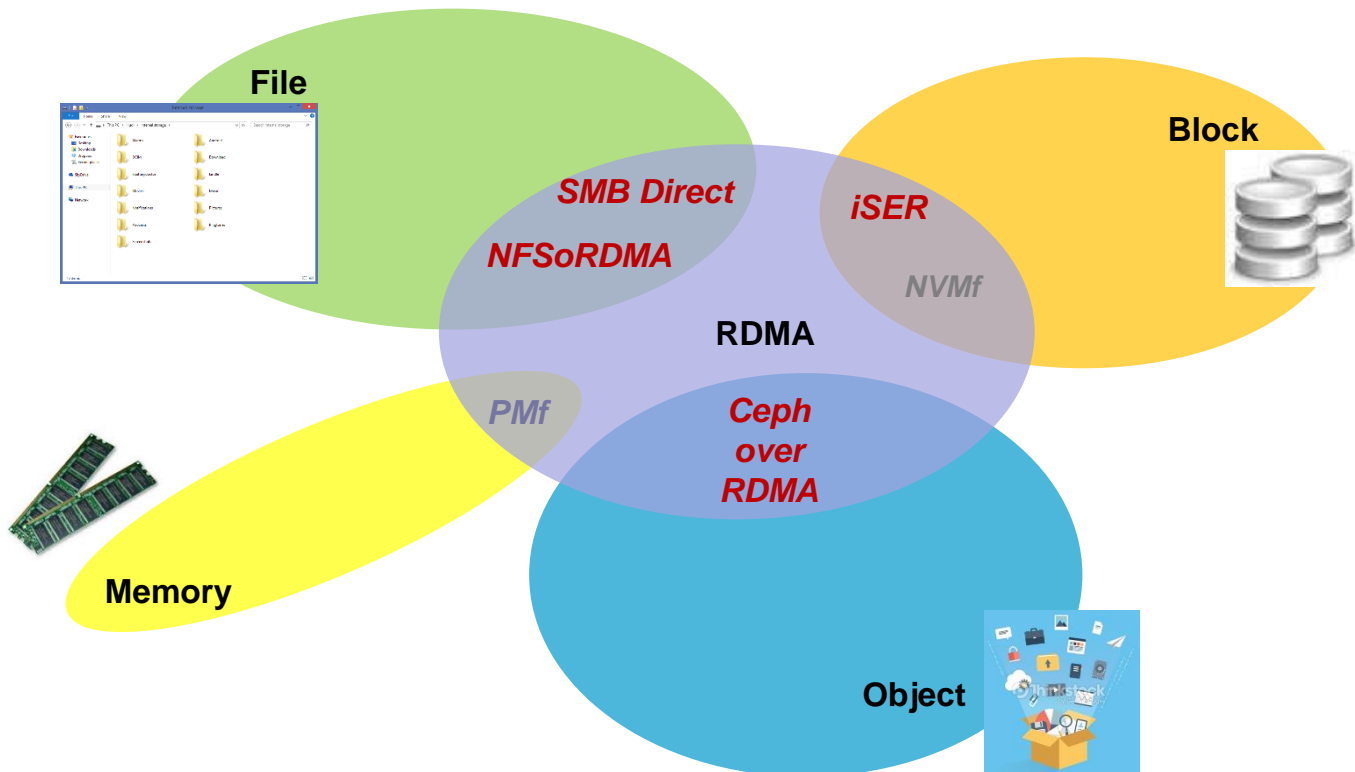
Open Source Flash Storage Solutions Performance

RDMA Storage Protocols

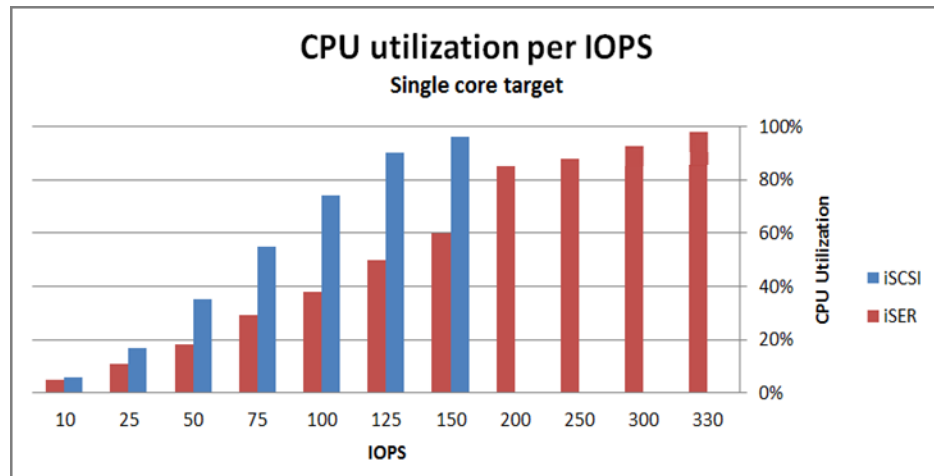
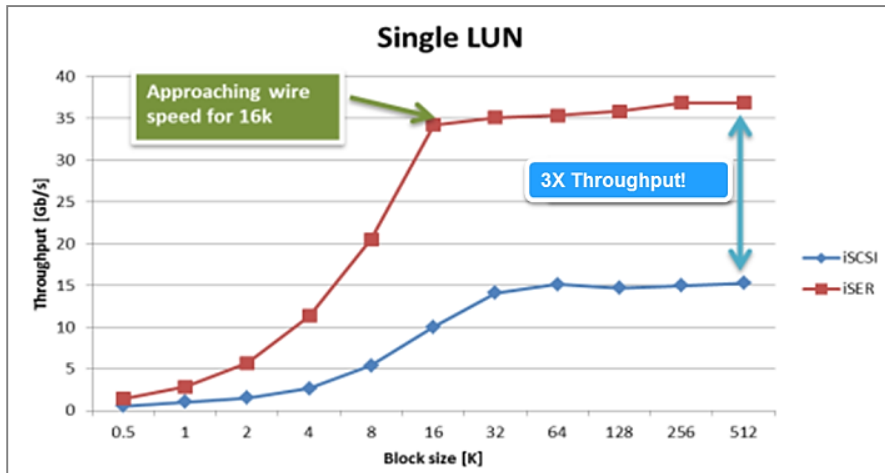
- [iSER](#)
- [SMB Direct](#)
- [Ceph over RDMA](#)

Non-Volatile Memory (NVM) Storage Protocols

- NVMe over Fabrics (NVMf)
- PMf (3D-XPoint)



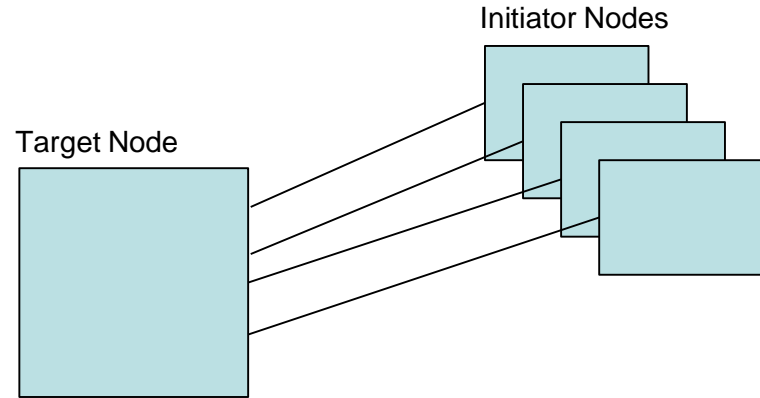
iSER Ethernet RDMA Block Performance



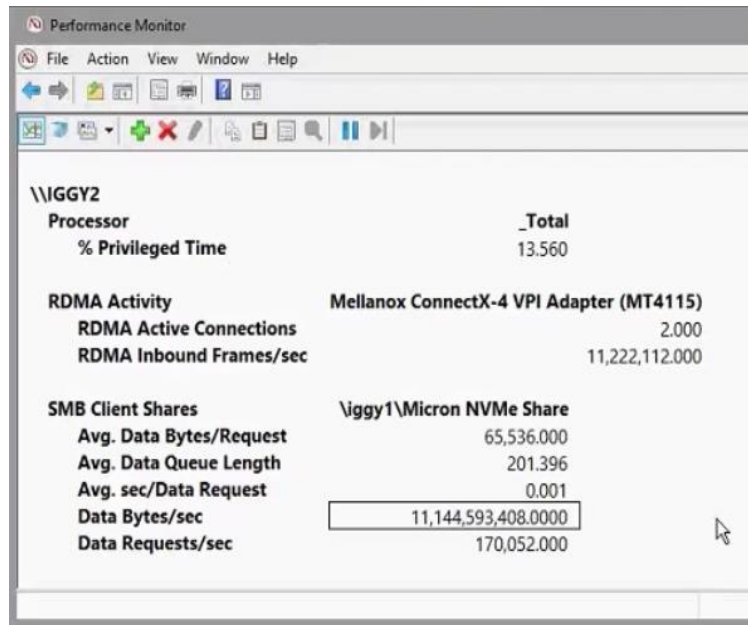
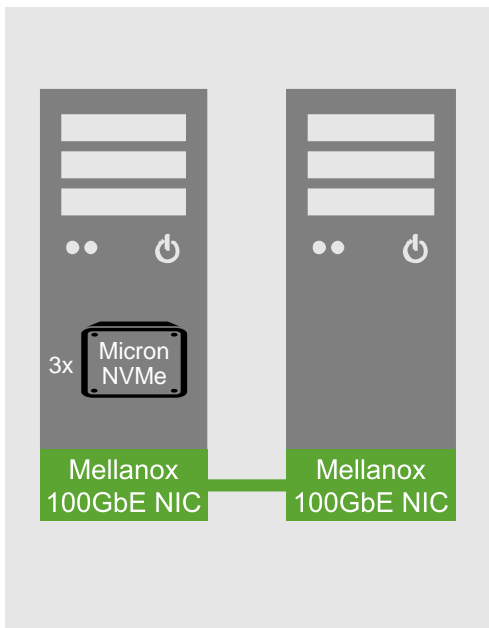
Higher Bandwidth and IOPS with Less CPU Utilization than iSCSI

iSER Performance Demo at FMS 2015

- Target node
 - Dual-socket x86 server
 - 4x40GbE NICs
 - iSER LIO target
 - 20xPM953 NVMe drives
- Initiators
 - Dual-socket x86 server
 - 1x40GbE NIC
- Performance
 - 2.1M – 4K Random Read
 - 17.2GB/s – 128K Seq Read

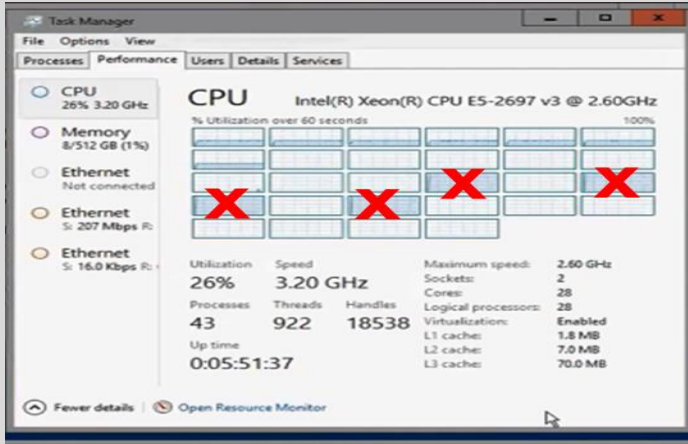


SMB Ethernet RDMA File Performance



Demo highlights:

- SMB3 using 100Gbps RDMA
- Storage Spaces using NVMe SSDs
- Over 11GB/sec over one NIC port
- 1ms latency with SMB3 storage
- Less than 15% CPU utilization



2X Better Bandwidth

Half the Latency

33% Lower CPU

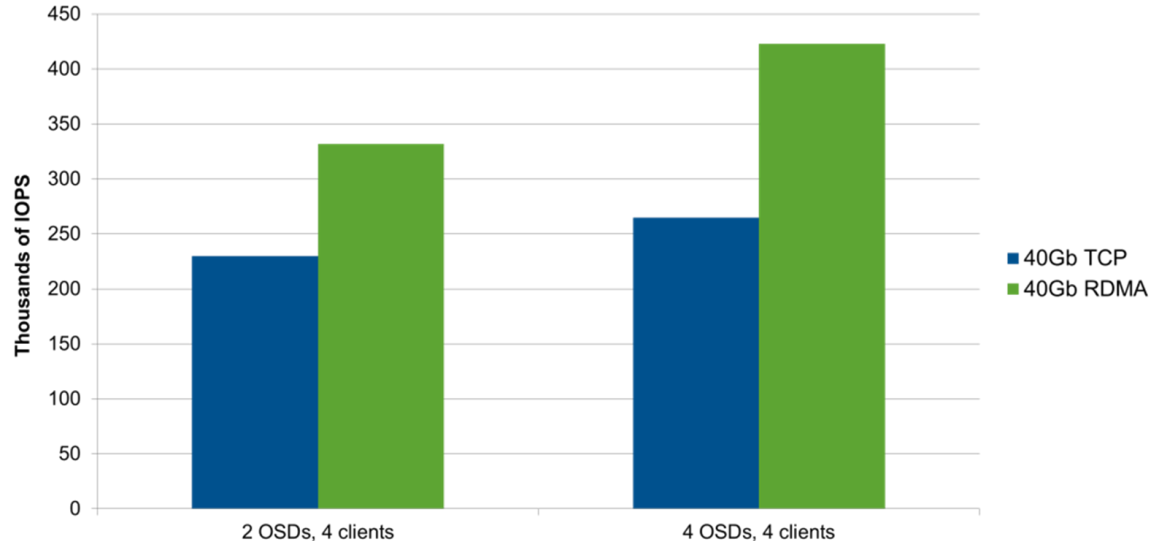
See the demo: <https://www.youtube.com/watch?v=u8ZYhUjSUoI>



- Without RDMA
 - 5.7 GB/s throughput
 - 20-26% CPU utilization
 - 4 cores 100% consumed by moving data
- With Hardware RDMA
 - 11.1 GB/s throughput at half the latency
 - 13-14% CPU utilization
 - More CPU power for applications, better ROI

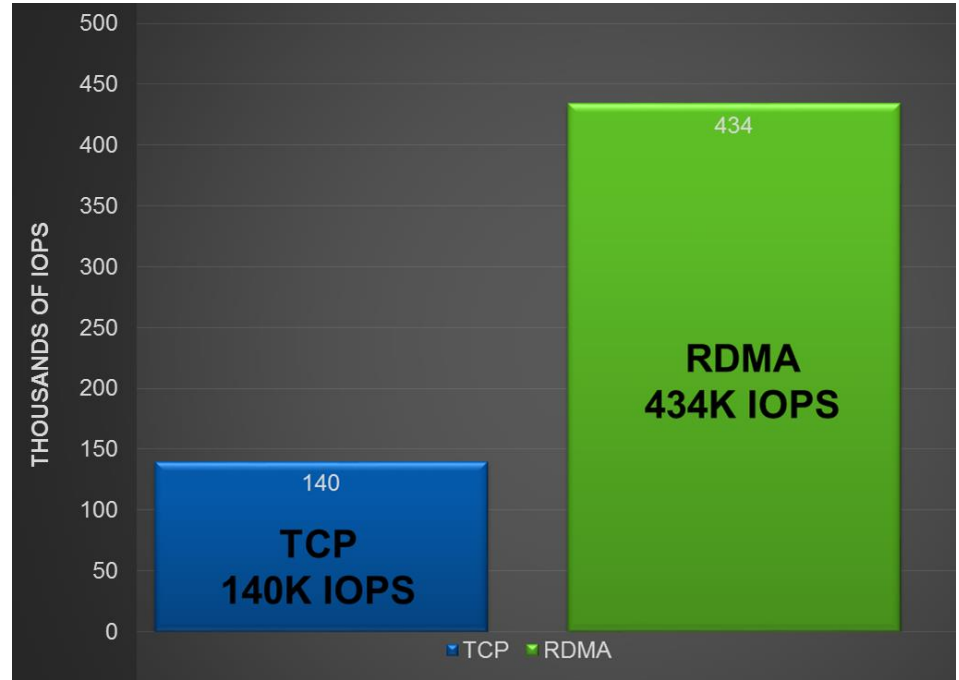
Object Storage with RDMA on Ceph

- RDMA is implemented in Ceph *Hammer Release* as Beta
- Tests show performance 30-40% better with RDMA

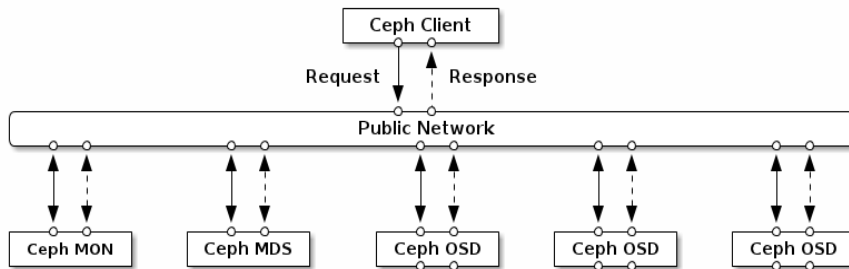


Ceph Performance with RDMA

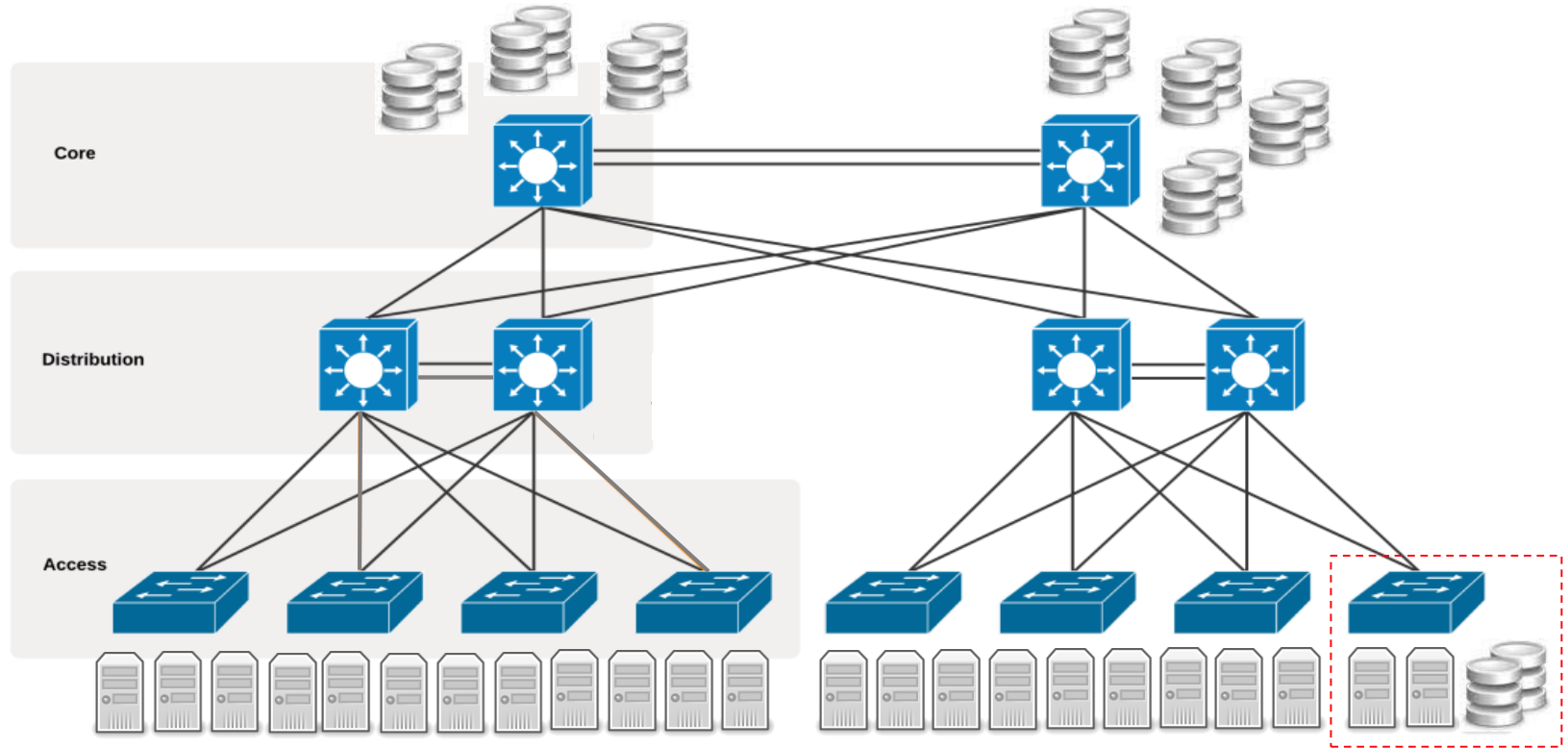
- Performance varies with work load
- Ceph Read IOPS: TCP vs. RDMA
 - High IOPS workload
 - Block sizes 32KB
- RDMA more than triples the performance
 - Less than 10usec latency under load



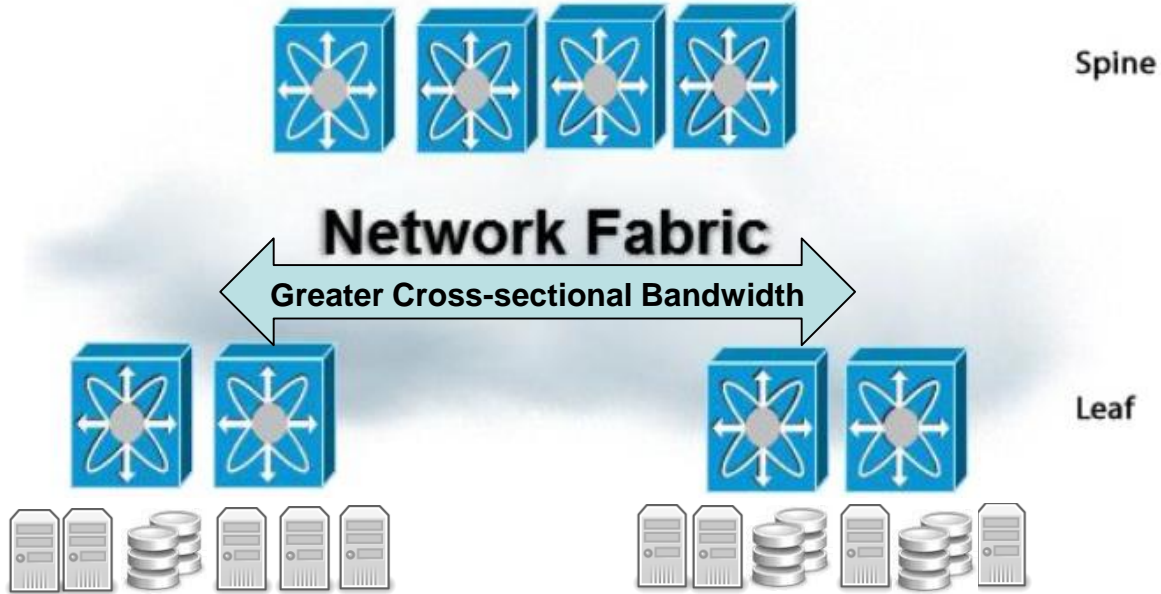
Flexible Ceph Architecture Makes RDMA Easy to Implement



RDMA Block Storage Architecture

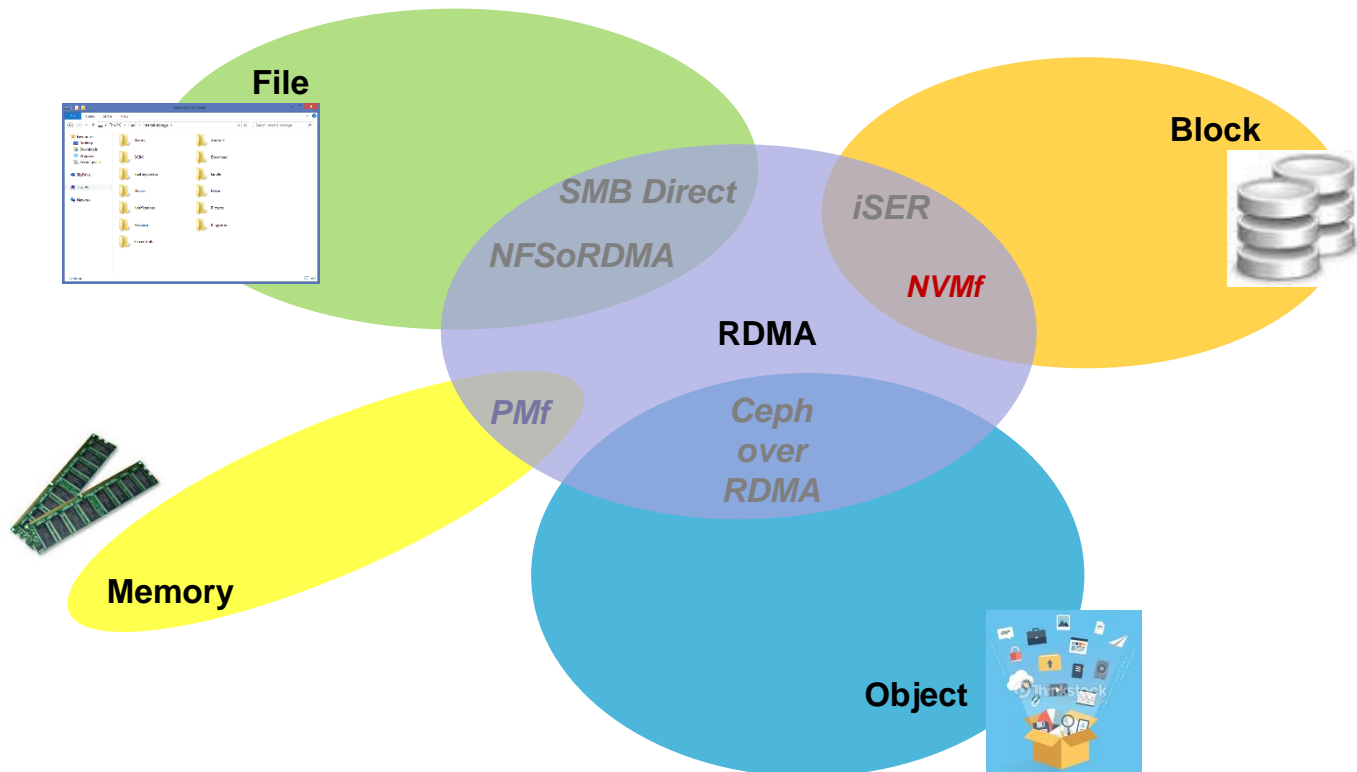


Leaf-Spine Architecture

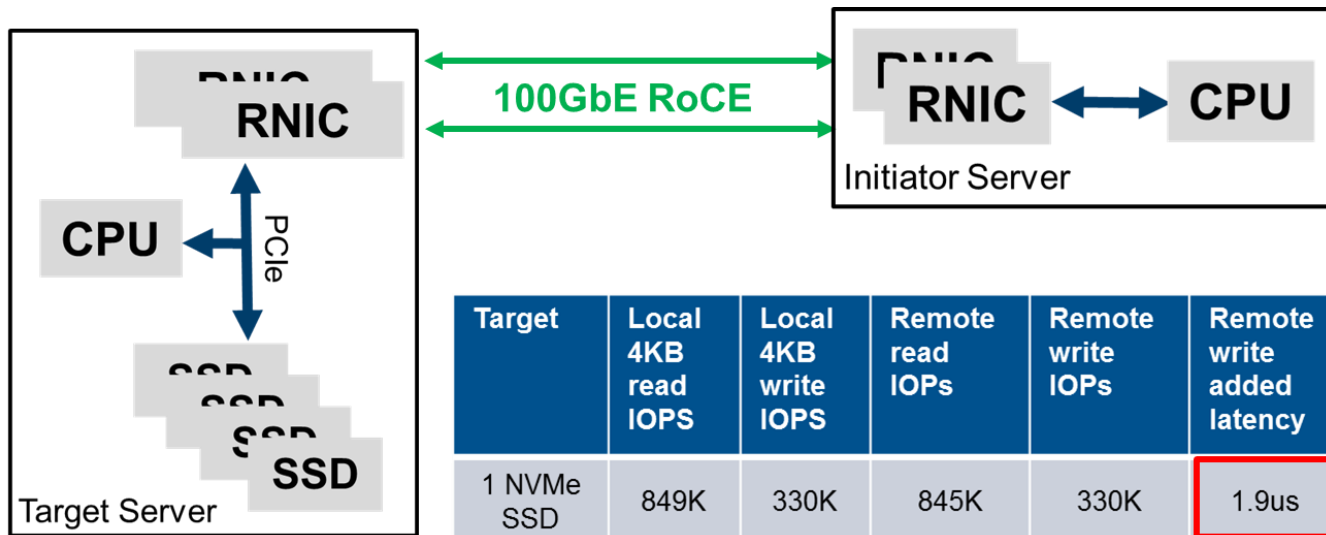


Open Source Flash Storage Solutions Performance

- RDMA Storage Protocols
 - iSER
 - SMB Direct
 - Ceph over RDMA
- Non-Volatile Memory (NVM) Storage Protocols
 - NVMe over Fabrics (NVMf)
 - PMf (3D-XPoint)



NVMe over Fabrics (NVMeF) Performance – FMS 2015



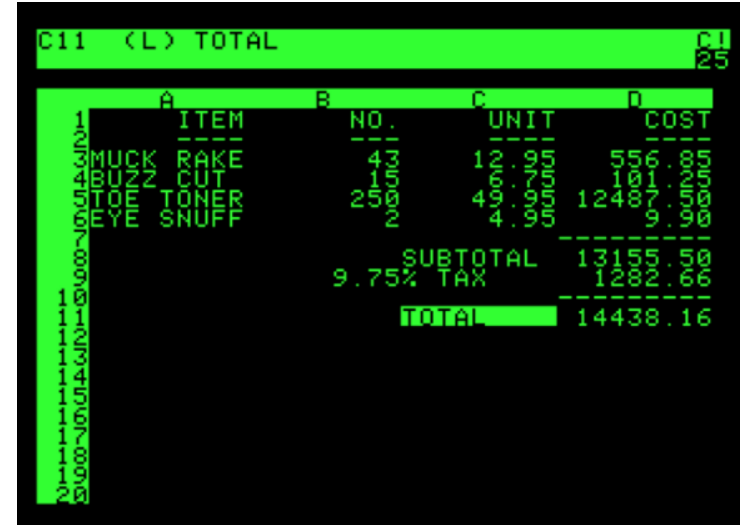
Target	Local 4KB read IOPS	Local 4KB write IOPS	Remote read IOPS	Remote write IOPS	Remote write added latency	Remote read added latency
1 NVMe SSD	849K	330K	845K	330K	1.9us	4.76us
4 NVMe SSDs	3406K	1333K	3388K	1332K	N/A	N/A

“Killer App”

From Wikipedia:

“In marketing terminology, a killer application (commonly shortened to killer app) is any computer program that is so necessary or desirable that it proves the core value of some larger technology...”

“One of the first examples of a killer app is generally agreed to be the VisiCalc spreadsheet...people spent \$100 for the software, then \$2000 to \$10,000 on the Apple computer they needed to run it”



A	B	C	D
ITEM	NO.	UNIT	COST
MUCK RAKE	43	12.95	556.85
BUZZ CUT	1	101.00	101.00
TOE TONER	25	49.95	1248.75
EYE SNUFF	2	4.95	9.90
		SUBTOTAL	13155.50
		9.75% TAX	1282.66
		TOTAL	14438.16



Apple Computer

NVMf Version 1.0 Open Source



Home Groups ProjectView

[Workspace](#) > [All Groups](#) > [My Groups](#) > Working Group - Fabrics Linux Driver

Working Group - Fabrics Linux Driver

Group Info
Group Chair: Bob Beauchamp, EMC

Group Email Addresses
Post message: fabrics_linux_driver@nvmexpress.org
Contact chair: fabrics_linux_driver-chair@nvmexpress.org

▪ **Related:**

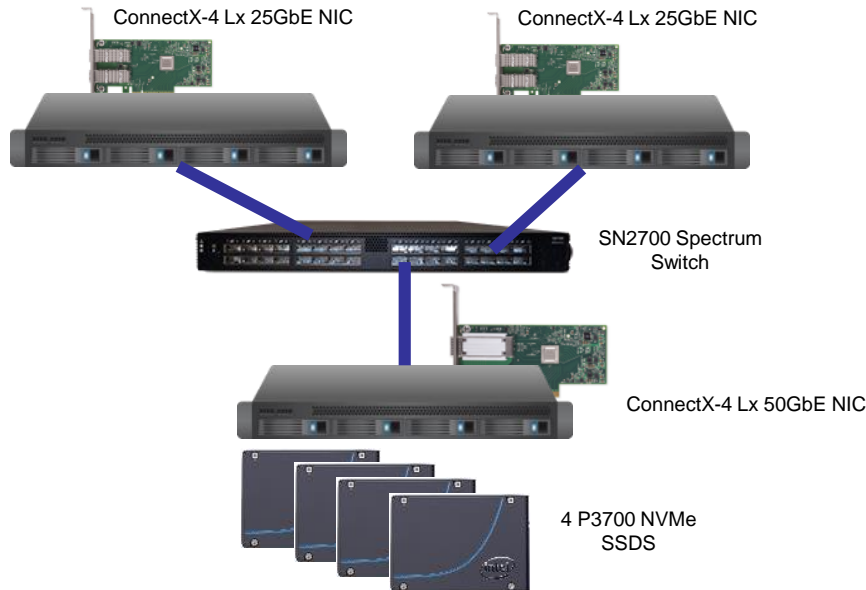
- Spec group (1st release by EOY)

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.

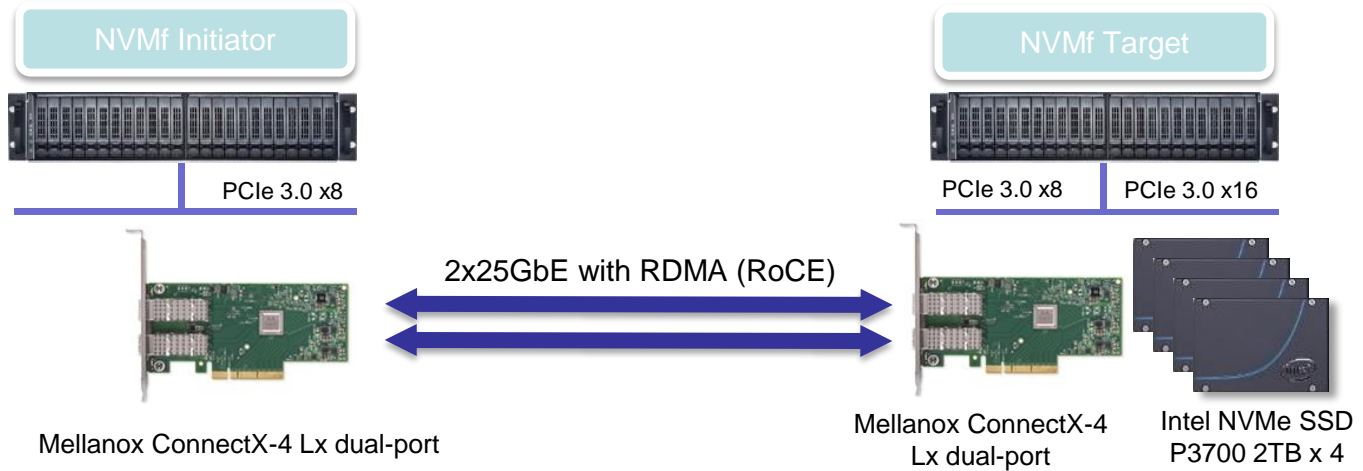
Performance with Community Driver

- Topology –
 - Two compute nodes
 - ConnectX4-LX 25Gbps port
 - One storage node
 - ConnectX4-LX 50Gbps port
 - 4 X Intel NVMe devices (P3700/750 series)
 - Nodes connected through switch



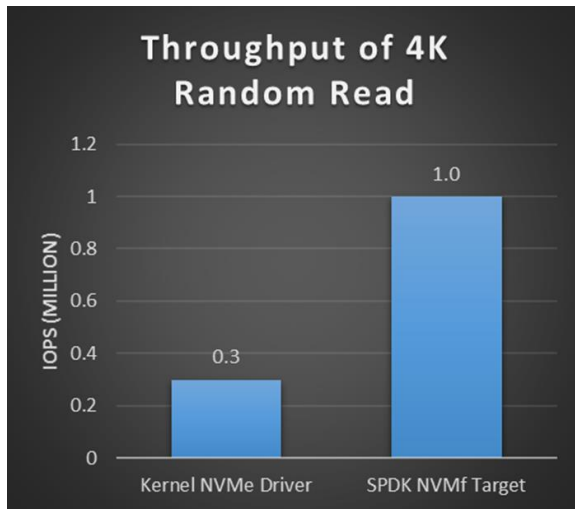
Added fabric latency				
~12us				
	Bandwidth (target)	IOPS (target)	# online cores	Each core utilization
BS = 4KB, 16 jobs, IO depth = 64	5.2GB/sec	1.3M	4	50%

SPDK NVMe Over Fabrics

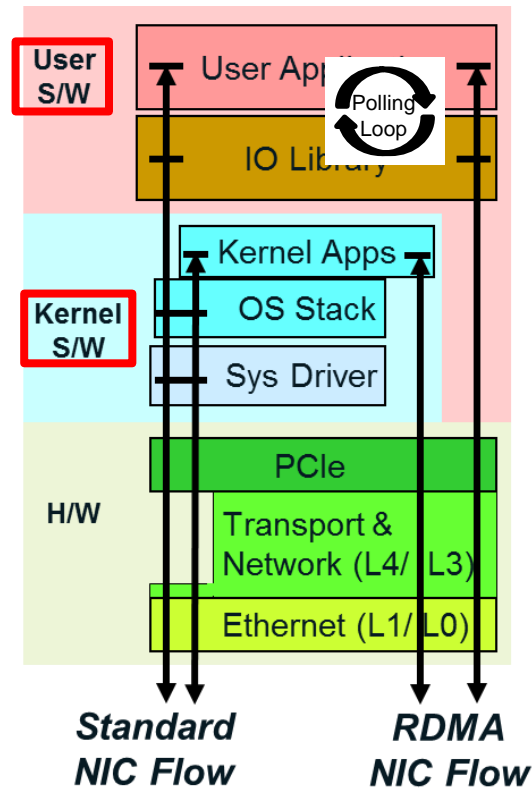


- SPDK software with user-space target and polling mode
- Mellanox ConnectX-4 Lx NICs, running RoCE
- 4 Intel NVMe P3700 SSDs connected to shared PCIe Gen3x16 bus
- Intel Xeon-D 1567 CPU on target side; Intel Xeon E5 2600 V3 on initiator side

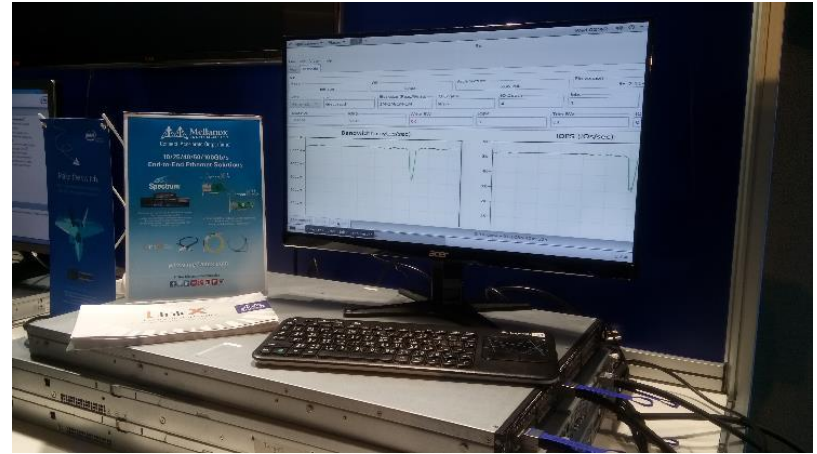
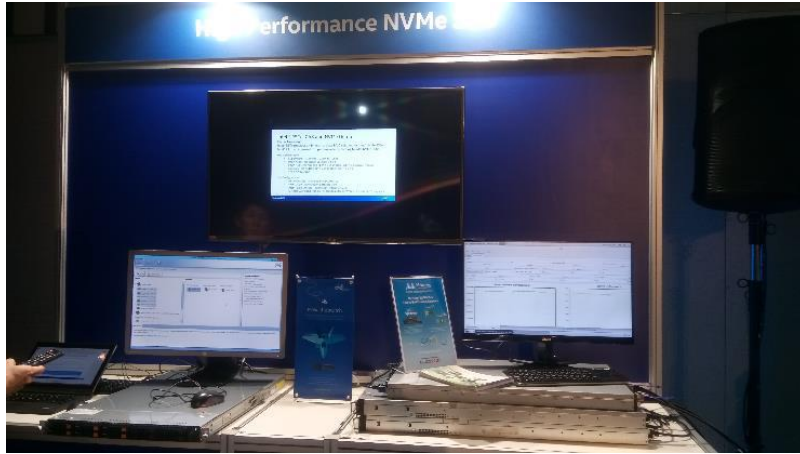
Intel SPDK NVMe Over Fabrics Performance



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



Intel NVMf Demo at Computex 2016 in Taiwan



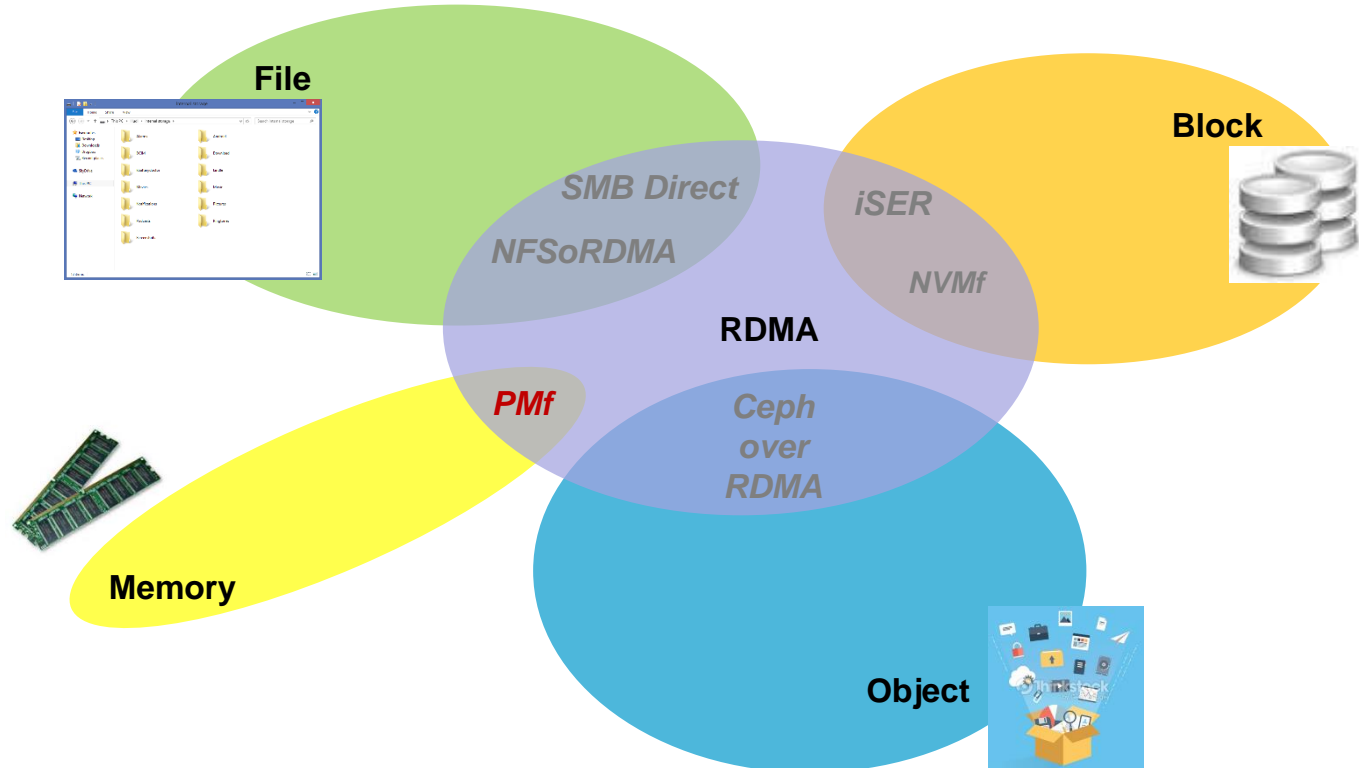
- ConnectX-4 Lx 1x50GbE NIC with RoCE
- Spectrum SN2100 25/40/50/100GbE switch (ports configured for 50GbE)
- Libaio engine, FIO benchmark tool, 4 jobs, IO Depth of 4
- 5.43 GByte/sec and 5304 IOPS with large read (1MB I/O size)

Open Source RDMA Software

- http://linux-iscsi.org/wiki/ISCSI_Extensions_for_RDMA
- <http://docs.ceph.com/docs/master/releases/>
- <https://www.samba.org/> (Windows SMB)
- <https://www.kernel.org/doc/Documentation/filesystems/nfs/nfs-rdma.txt>
- <git://git.infradead.org/nvme-fabrics.git>
- <http://www.spdk.io/spdk/doc/nvmf.html>
- <https://github.com/SoftRoCE>
- <https://community.mellanox.com/docs/DOC-2283>

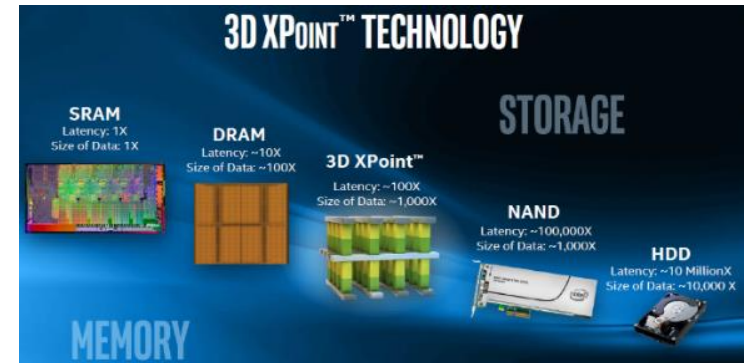
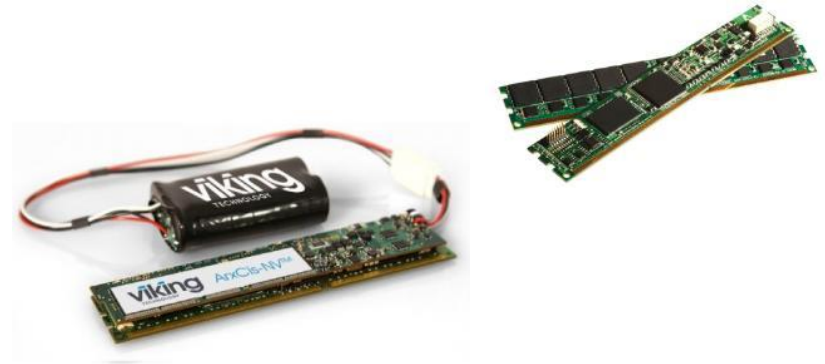
Disruptive Technology - Persistent Memory in Storage

- RDMA Storage Protocols
 - iSER
 - SMB Direct
 - Ceph over RDMA
- Non-Volatile Memory (NVM) Storage Protocols
 - NVMe over Fabrics (NVMf)
 - PMf (3D-XPoint)

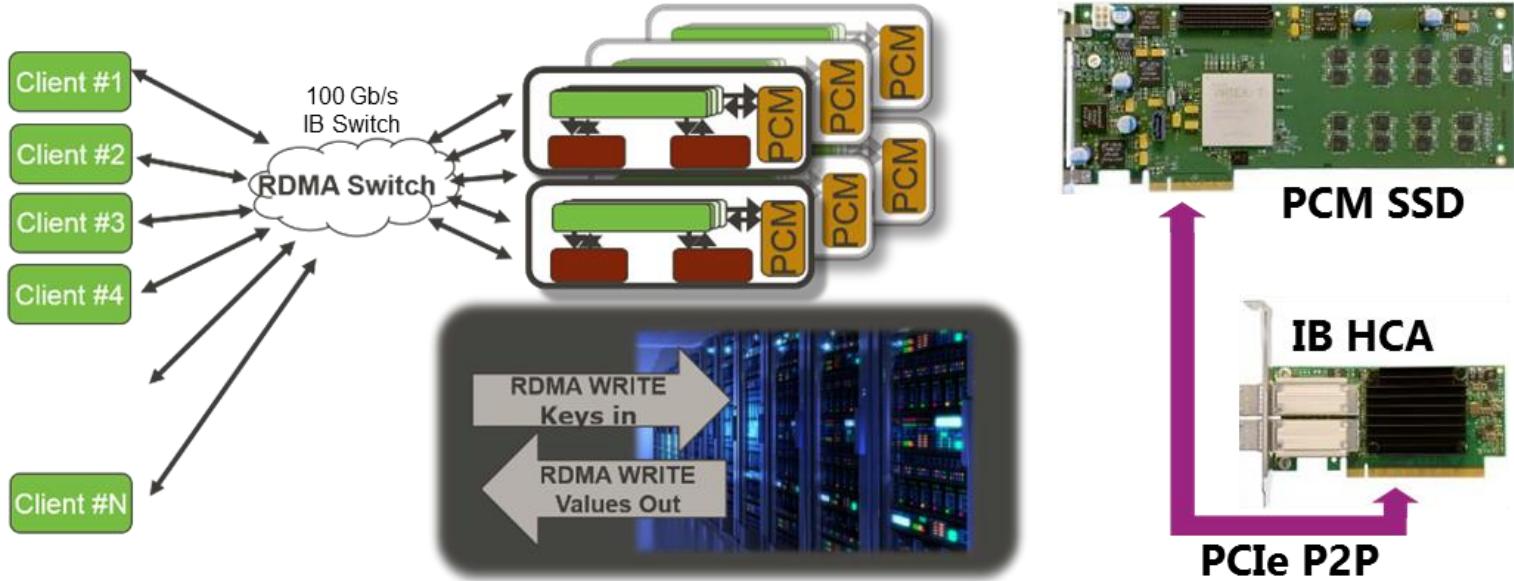


Persistent Memory in Storage

- Storage with memory performance
 - Large Write Latency Improvements over Flash
 - Byte Addressability
 - E.g. 3dxcpoint, NVDIMM, NVRAM, RERAM
- Emerging Eco-system for Direct Attach Storage
 - SNIA NVM Programming Model TWIG
 - Memory mapping of the storage media
 - PMEM.IO, DAX changes in file system stack
- Next step is Remote Access
 - SNIA NVM PM Remote Access for High Availability

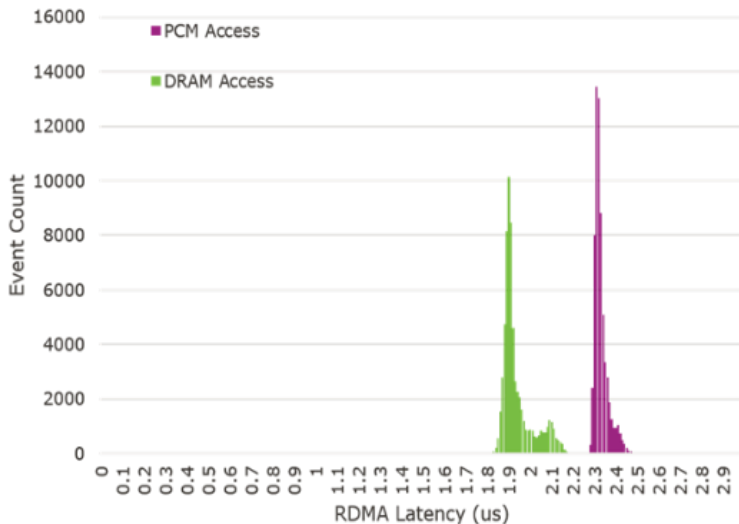


HGST FMS 2015 Demo

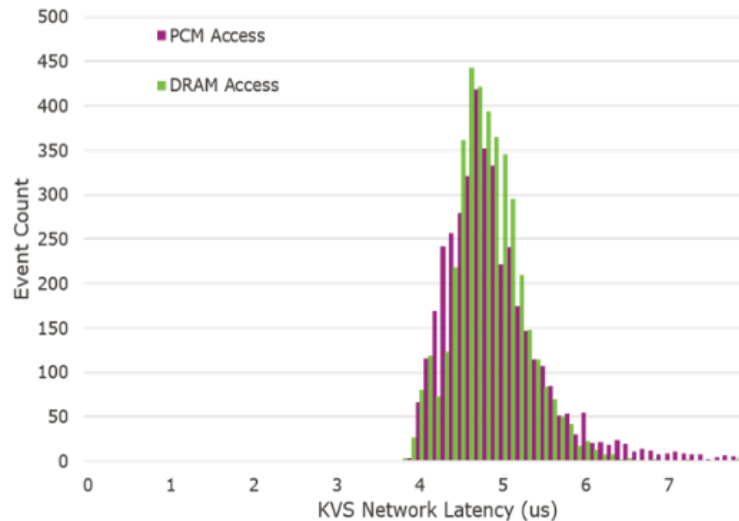


Equivalent Application Performance

PCM Hardware Latency is only 18% Slower than DRAM...



...and has Equivalent KVS Application Performance!



<https://www.hgst.com/company/media-room/press-releases/HGST-to-Demo-InMemory-Flash-Fabric-and-Lead-Discussions>

Conclusions

- With Flash and faster network speeds Block, File, and Object storage application performance improves
- By adding RoCE RDMA network technology performance can be enhanced dramatically and you can future proof your network for next generation storage
- The software that powers RDMA technology is available through open source



Thanks!

robd@mellanox.com