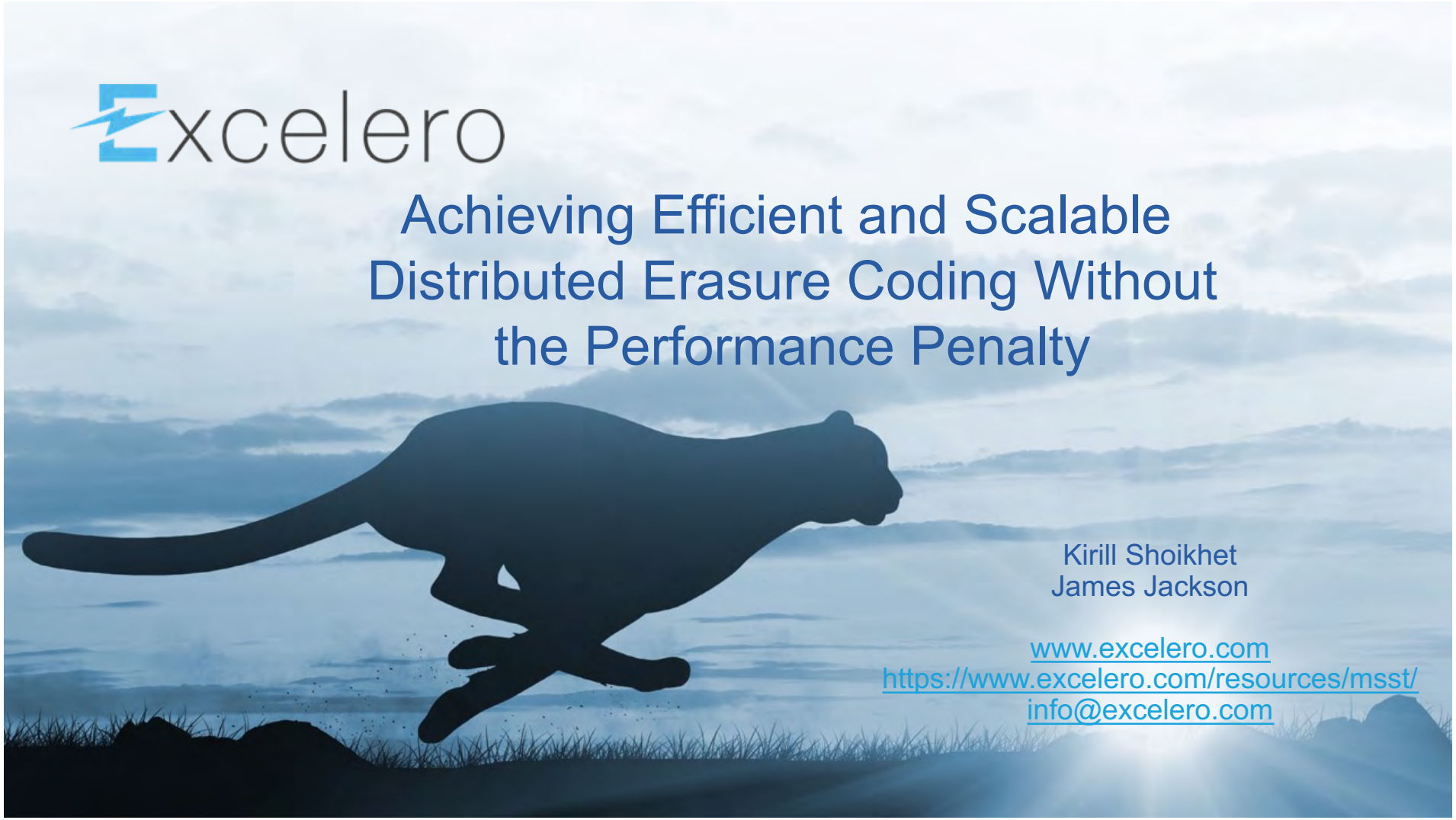

Excelero
Achieving Efficient and Scalable Distributed Erasure Coding Without the Performance Penalty
Kirill Shoikhet
James Jackson
www.excelero.com
https://www.excelero.com/resources/msst/
info@excelero.com

FINALISTS

DataCore SOFTWARE
elastifile Cross-Cloud Data Fabric
Excelero
HEDVIG
IBM Spectrum Storage
Pivot
Qumulo
SCALITY
openSUSE
WEKA.IO
RED HAT CEPH STORAGE

Excelero

Excelero's NVMesh® Named 2017 Product of the Year in Software-Defined Storage

by Storage Magazine and SearchStorage

Excelero

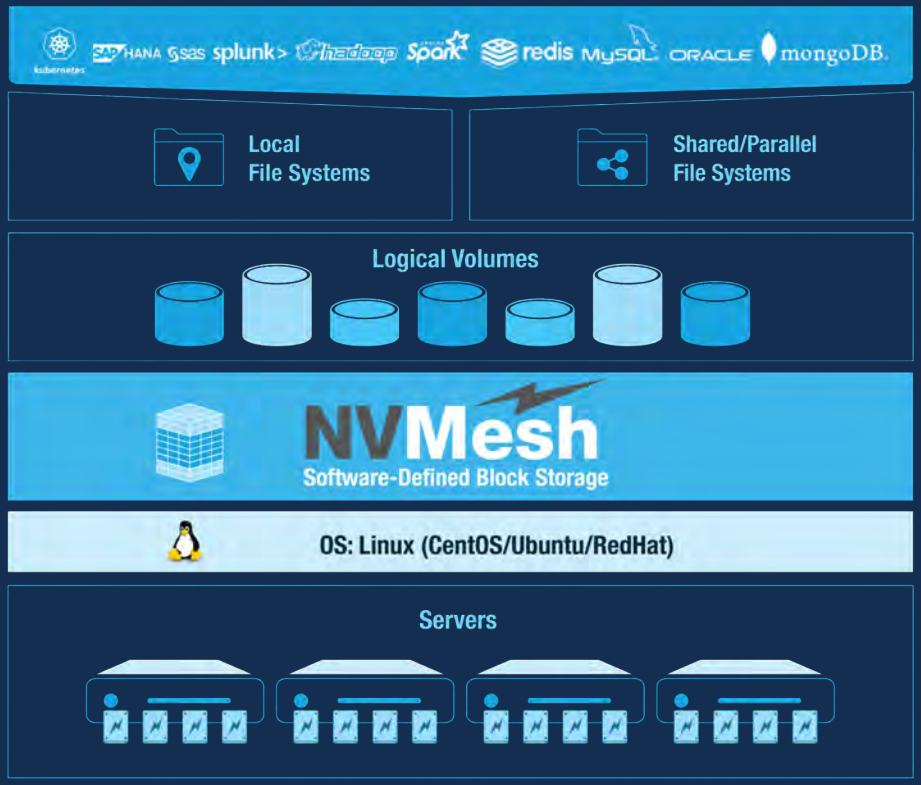
Named a 2018 Cool Vendor in Storage Technologies by Gartner

What is NVMesh server SAN?

NVMesh allows unmodified applications to utilize pooled NVMe storage devices across a network at local speeds and latencies.

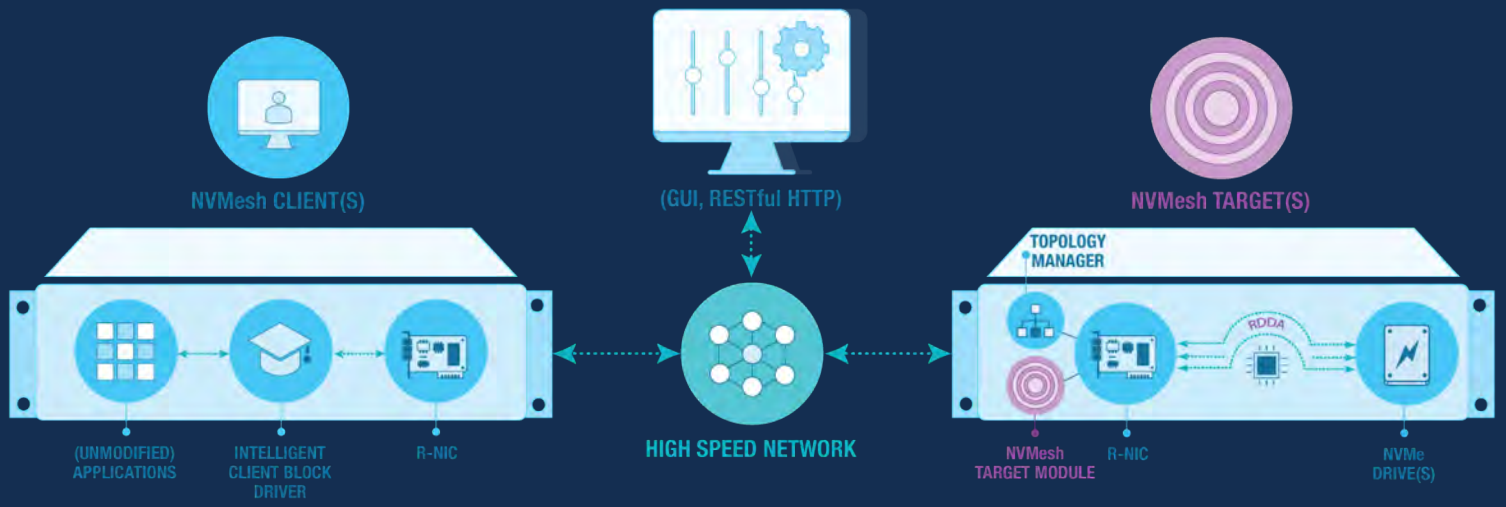
Distributed NVMe storage resources are pooled with the ability to create arbitrary, dynamic block volumes that can be utilized by any host running the NVMesh block client.

You can access any of your storage in the network as (shared) volumes from anywhere in the network.



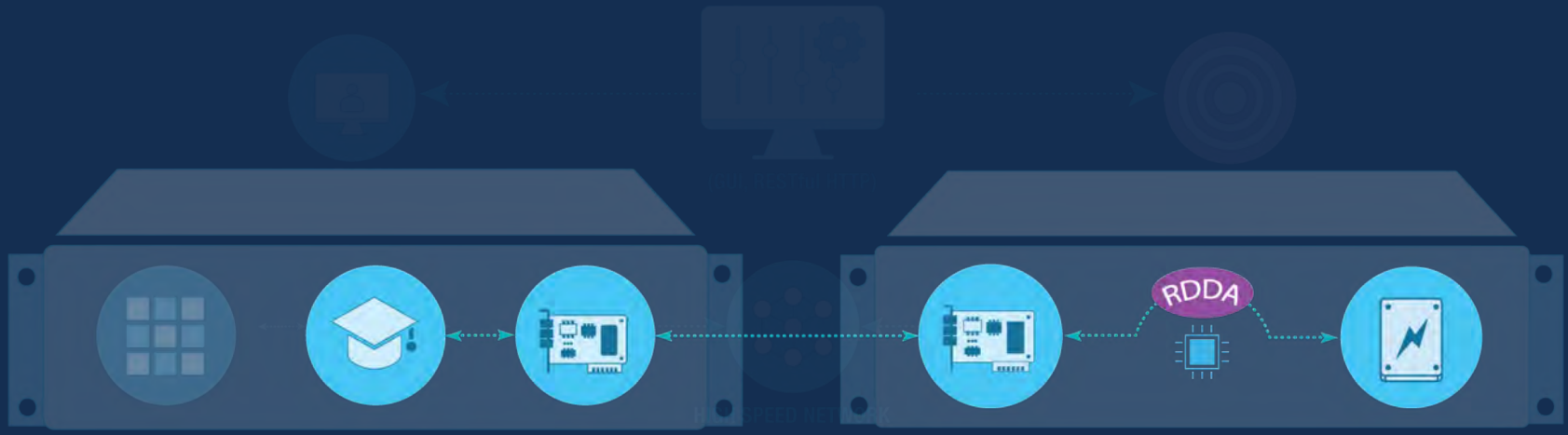
Excelero

NVMesh Software Components



Excelero

RDDA



RDDA (Remote Direct Device Access) uses RDMA to access target NVMe drives. **ZERO** target CPU usage.

Benefits of client-side architectures

- NVMe lockless, reduces software stack
- NVMeoF maps onto RDMA fabrics like IB or RoCE
- Storage controller intermediary provides features (cache, device control, redundancy, host sharing)
  - NVMe drives saturate host bandwidth, no need for storage controller
    1. Build HW accelerators to offload functionality from CPUs
    2. Remove intermediaries by moving functionality to client side
       - Much larger client compute pool

This presentation focuses on the client-side approach.

Trends towards larger writes

- Many existing applications use block storage for persistence
  - Write optimization is for bandwidth
  - Read optimization is for latency
  - Transaction size 100's of KB to MB
  - Example: default Luster size 4MB
- In-memory processing popular with rendering, analytics, etc.
  - Bandwidth highest priority, size at least 128 KB
  - IBM Spectrum Scale burst buffer one example
- *Movement to large, aligned I/Os*

Models for Comparison and Analysis

Assume mixed-use NVMe: 200K 4K IOPs random write, 800K 4K random read

Converged Model:

- Compute nodes *also* storage nodes
- Standard servers
  - Four NVMe drives
  - Two 25 Gb/s RNICs
  - 10 servers

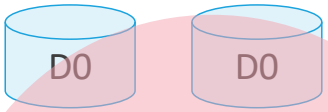
We'll focus on this approach

Disaggregated Model:

- Separate "target" storage nodes and "client" compute nodes
- Targets JBOF or dedicated servers with many (e.g. 24) NVMe
  - RNICs up to 8 100Gb/s ports
- Clients local boot storage & two 25 Gb/s RNICs

Protect Your Data: Replication (Mirroring)

40 disks, 10 hosts
Disks read 3.2GB/s, write .8GB/s
Host read and write 6 GB/sec



2-way replication:

- Storage efficiency 50%
- Tolerates 1 drive failure
- Write 16 GB/s application data
- 32 GB/s Network data
- Read 60 GB/s (network limited)

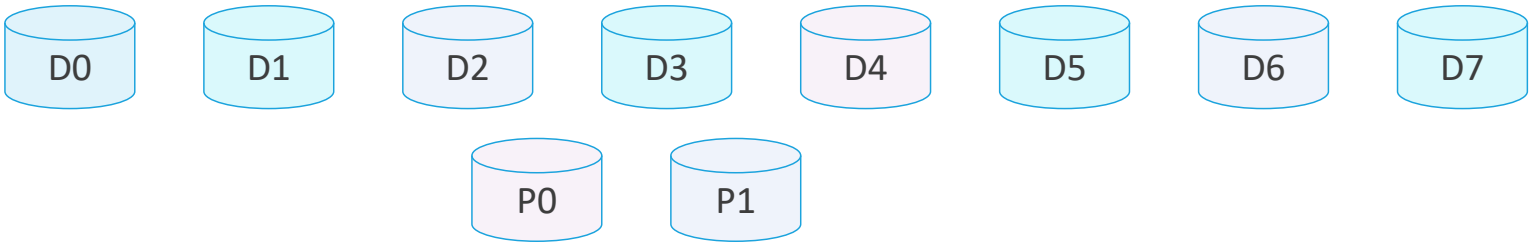
3-way replication:

- Storage efficiency 33%
- Tolerates 2 drive failures
- Write 10 GB/s application data
- 30 GB/s network data
- Read 60 GB/s (network limited)

Protect Your Data: Erasure Coding (RAID6)

Each stripe has N data blocks and P "parity" blocks (e.g., 8+2)

- Storage efficiency is $^{N}/_{(N+P)}$ (80% for 8+2)
- Higher N means slower writes for partial stripe
- Larger writes can mitigate or eliminate partial stripe writes
- Tolerates 2 drive failures
- Like mirroring, each drive should be on a different server



Erasure Coding Issues

- "Write hole"
  - Failure in the middle of multiple writes
  - Parity doesn't match stripe data
  - *Data corruption threat* – ignored by some software RAID 5/6 products
  - Journaling makes operation atomic and consistent

- Journal requires separate writes – journal must finish before starting data
  - Stripe write is 10 drives (8 + 2) → 32KB becomes 80 KB writes with journal
  - Journal is small; still 80% storage efficiency
  - Journal is on same physical drive as data – identical failure characteristics
  - Cuts write speed in half

Erasure Coding - Baseline

10 nodes each with 4 NVMe drives and two 25 Gb/sec RNICs (6GB/sec per node)
Aggregate 40 drives, and 60 GB/sec network speed

Baseline assumes NVMeoF; 8 + 2; 32 KB ***aligned*** writes

- Journal doubles the number of drive writes and network writes.
  - 32 KB application write becomes 80KB disk and network write

- Max Read = 10<nodes> * 6 GB/s<NICs> = 60 GB/s (network limited)
- Max Write = 40<drive> * .8 GB/s<drive> * .5<journal> = 16 GB/s disk data
  - 16 GB/s network data
  - 6.4 GB/s application data

NVMesh improves balance and efficiency

- RDDA transfers data on network *once* for journal and data
  - Journal doubles the number of drive writes but not network writes
  - 32 KB application write becomes 40 KB network write and 80 KB disk write

10 nodes with 4 NVMe drives and two 25 Gb/sec RNICs (6GB/sec per node)
Aggregate 40 drives, and 60 GB/sec network speed

- Max Read = 10<nodes> * 6 GB/s<NICs> = 60 GB/s (network limited)
- Disk Write = 40<drive> * .8 GB/s<drive> * .5<journal> = 16 GB/s disk data
  - 8 GB/s network data
  - 6.4 GB/s application data

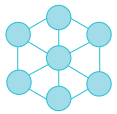
NVMe Persistent Memory Regions

- NVMesh allocates PMR across all clients, retained across failures

- Journal in PMR, so only *half* as many drive writes needed. PMR is part of the drive
  - Journal writes don't increase drive writes *or* network writes, because of PMR

10 nodes with 4 NVMe drives and two 25 Gb/sec RNICs (6GB/sec per node)
Aggregate 40 drives, and 60 GB/sec network speed

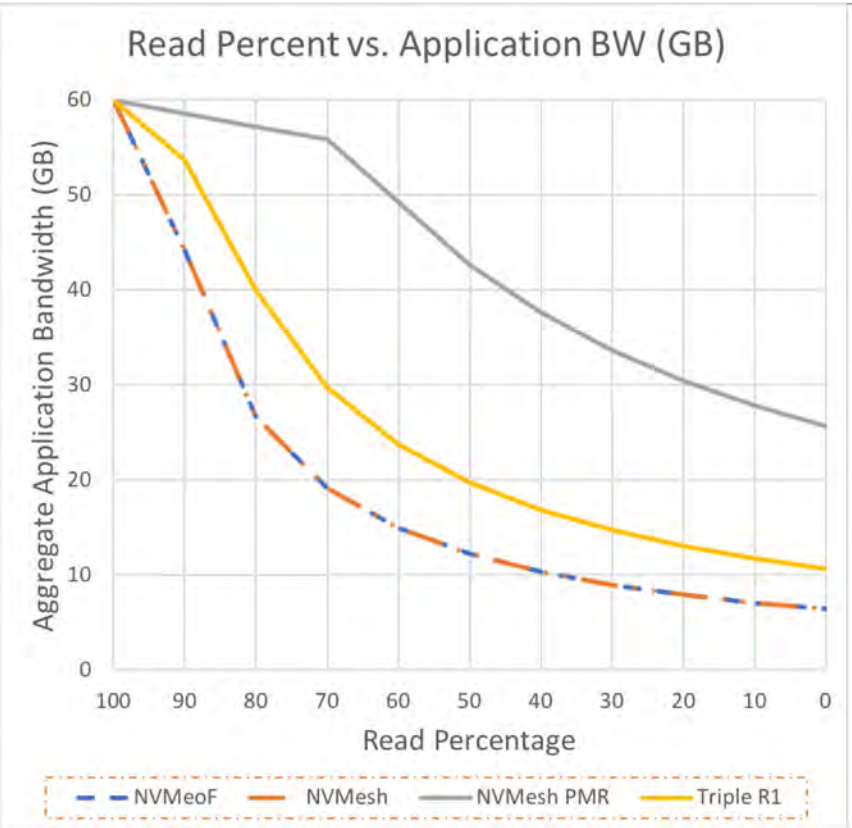
- Max Read = 10<nodes> * 6 GB/s<NICs> = 60 GB/s (network limited)
- Max Write = 40<drive> * .8<overhead> * .8 GB/s<drive> = 16 GB/s disk write
  - 16 GB/sec network data
  - 12.8 GB/sec application data

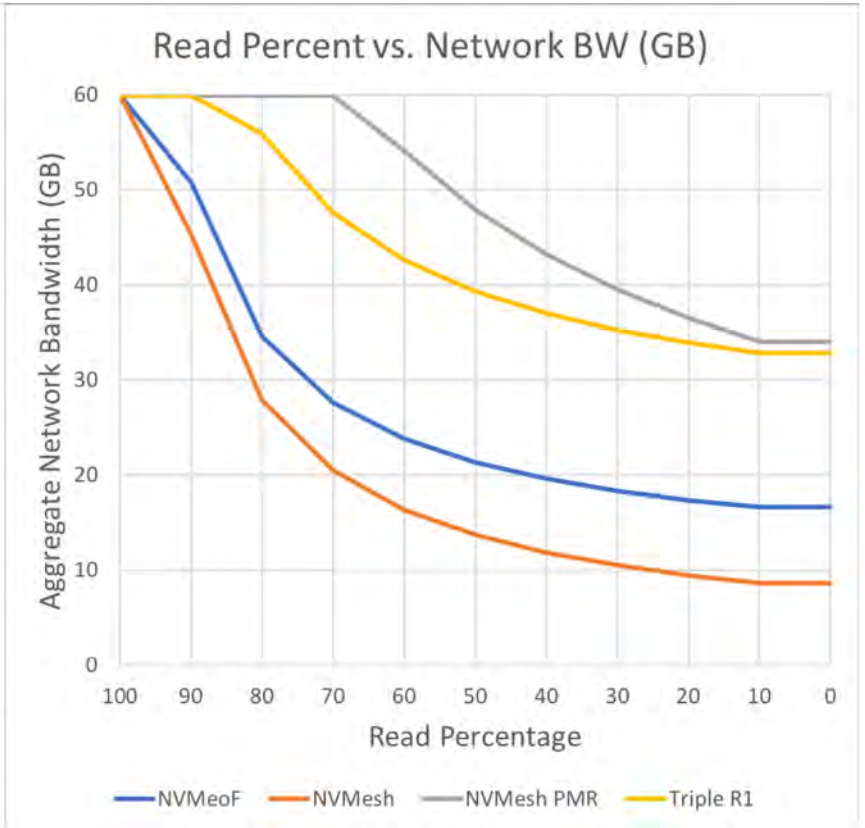


Comparison of the write methods

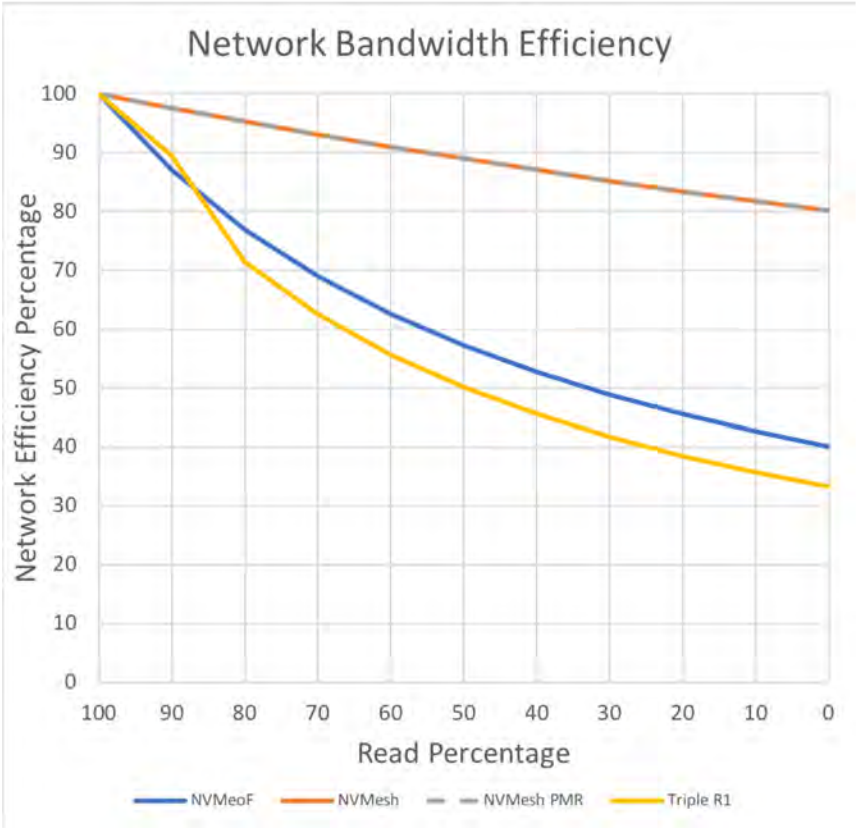
	NVMeoF	NVMesh	NVMesh PMR	Triple redundant
Space efficiency	.8	.8	.8	.33
Application BW	6.4	6.4	12.8	10
Network BW	16	8	16	30
Network efficiency	.4	.8	.8	.33

For Visual Learners





For Visual Learners (cont.)



Excelero


Excelero
Questions?