


A violin is the central focus, positioned diagonally from the bottom left towards the top right. The background is a vibrant, abstract composition of light waves and gradients in shades of green, yellow, and purple, creating a sense of motion and energy.

INNOR

Lustre on Flash for Modern Workloads

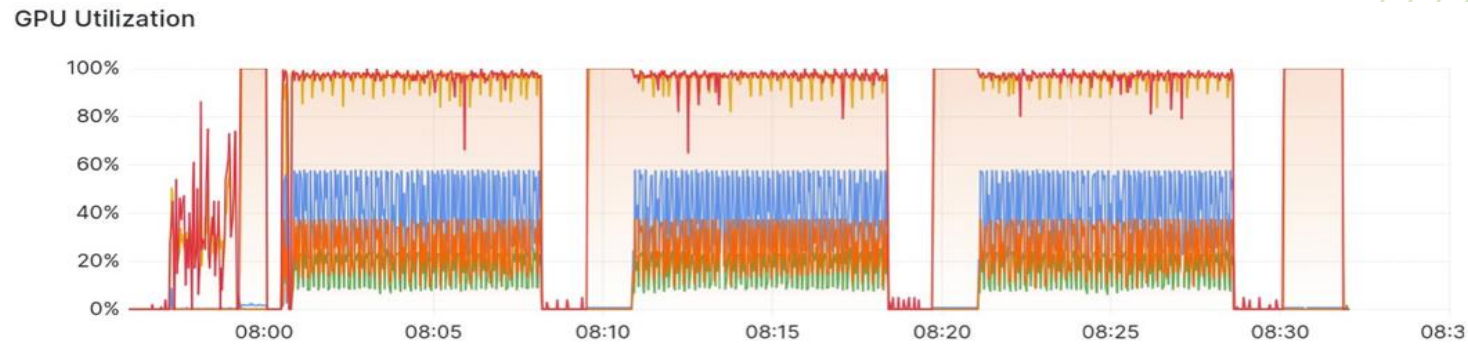
Daniel Landau, System Architect
Dr. Dmitry Livshits, CEO

Lustre is great for HPC,
but is it still relevant to
run AI workloads?

The background features a glowing, grid-like pattern that transitions from a dark red on the left to a bright green on the right. The grid is composed of small, interconnected nodes and lines, creating a sense of depth and movement. The overall effect is reminiscent of a digital or scientific visualization, possibly representing data flow or a network structure.

Performance requirements for AI

COMPUTE



I/O



More details: https://www.depts.ttu.edu/hpcc/events/LUG24/slides/Day2/LUG_2024_Talk_15-AI_Workload_Optimization_with_Lustre.pdf

AI Workloads Requirements for Lustre

Read Operations

Both random and sequential read performance is important (depending on AI models)

Write Operations

Equally important as read speed to minimize checkpoint time and GPU idle time

Storage Type

All Flash (or better, all NVMe) is needed to achieve the performance requirement

Data Integrity

Poor data integrity implementation can limit NVMe performance

xiRAID Classic Architecture

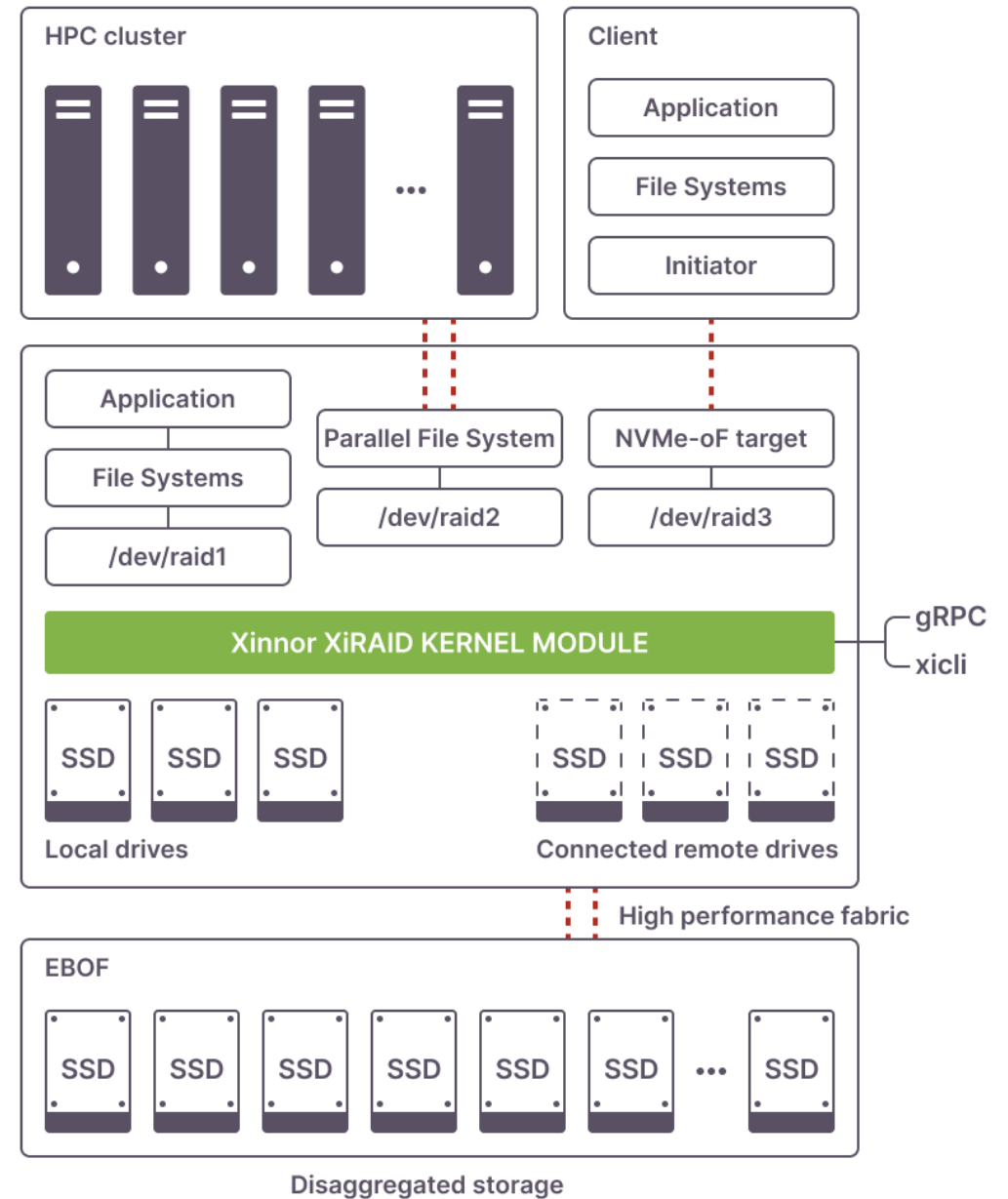
xiRAID is a software RAID made up of a Linux kernel module and a management utility (CLI)

✓ Installation by means of .rpm / .deb packages

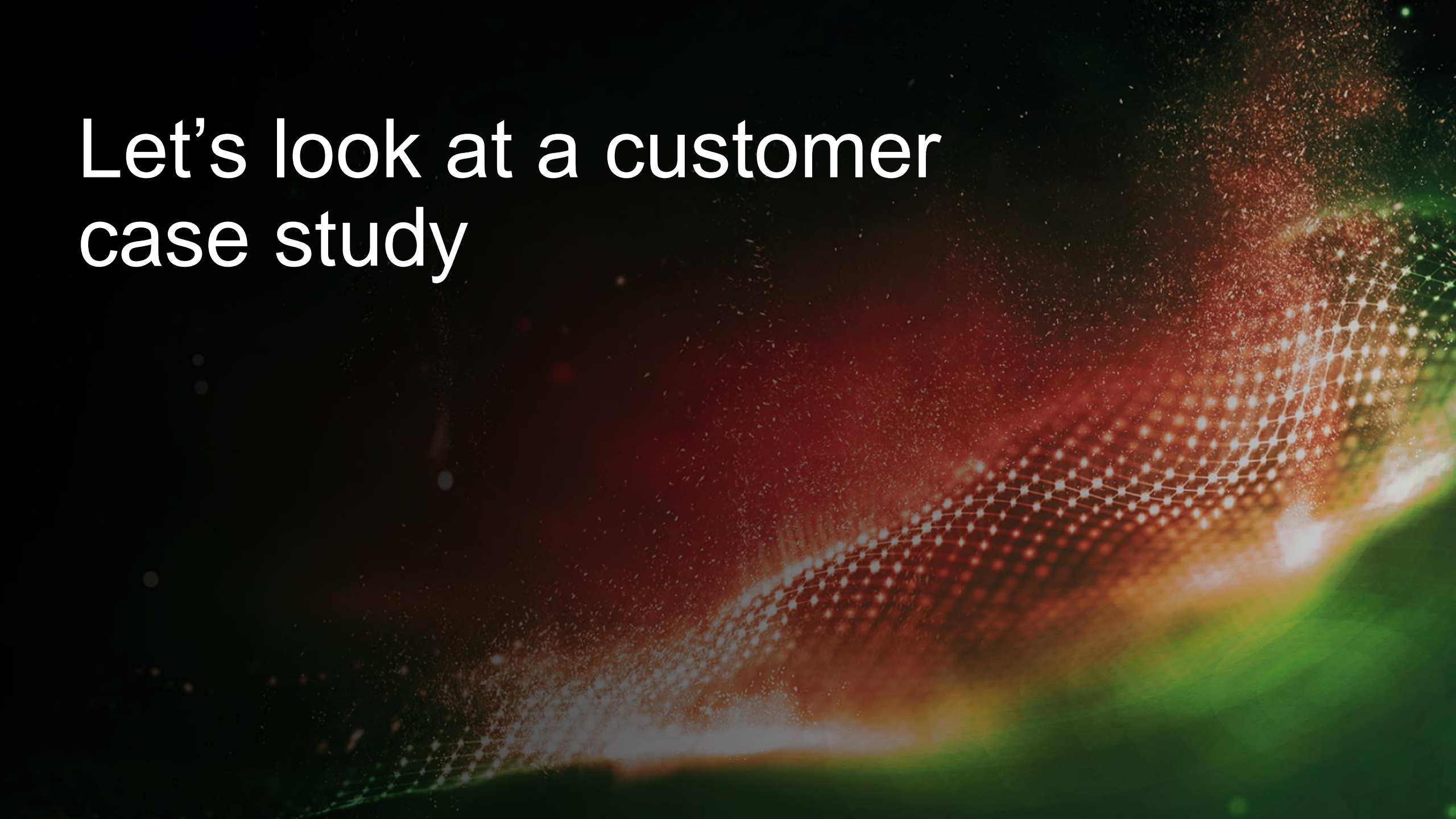
✓ Ready builds for the most popular Linux distributions:

- ✓ Oracle 8.4, 8.6, 9,
- ✓ RHEL 7.9, 8, 9.0, 9.1, 9.2, 9.3, 9.4
- ✓ Ubuntu 20.04, 22.04, 23.04
- ✓ AlmaLinux 9
- ✓ RockyLinux 9
- ✓ Proxmox 7.2

- ✓ Works with local and remote drives. Provides data availability for disaggregated storage environments
- ✓ Exports RAID as a standard Linux block device. No need to modify application stack



Let's look at a customer
case study

The background features a dynamic, abstract composition. A grid of glowing points, transitioning from red to green, curves across the right side of the frame. The overall aesthetic is futuristic and digital, with a dark space filled with fine particles and light trails.



FAU - Friedrich-Alexander-Universität Erlangen-Nürnberg

- GPU cluster "Alex": <https://doc.nhr.fau.de/clusters/alex/>
- This cluster is used for HPC, ML (millions of small files for ML) and AI

Customer challenge

- Ceph as storage solution was installed
- Performance was not sufficient (magnitude of order problems)
- 3-way Ceph replication means significant loss in used capacity
- Alternative solution was needed in short term perspective

in cooperation with

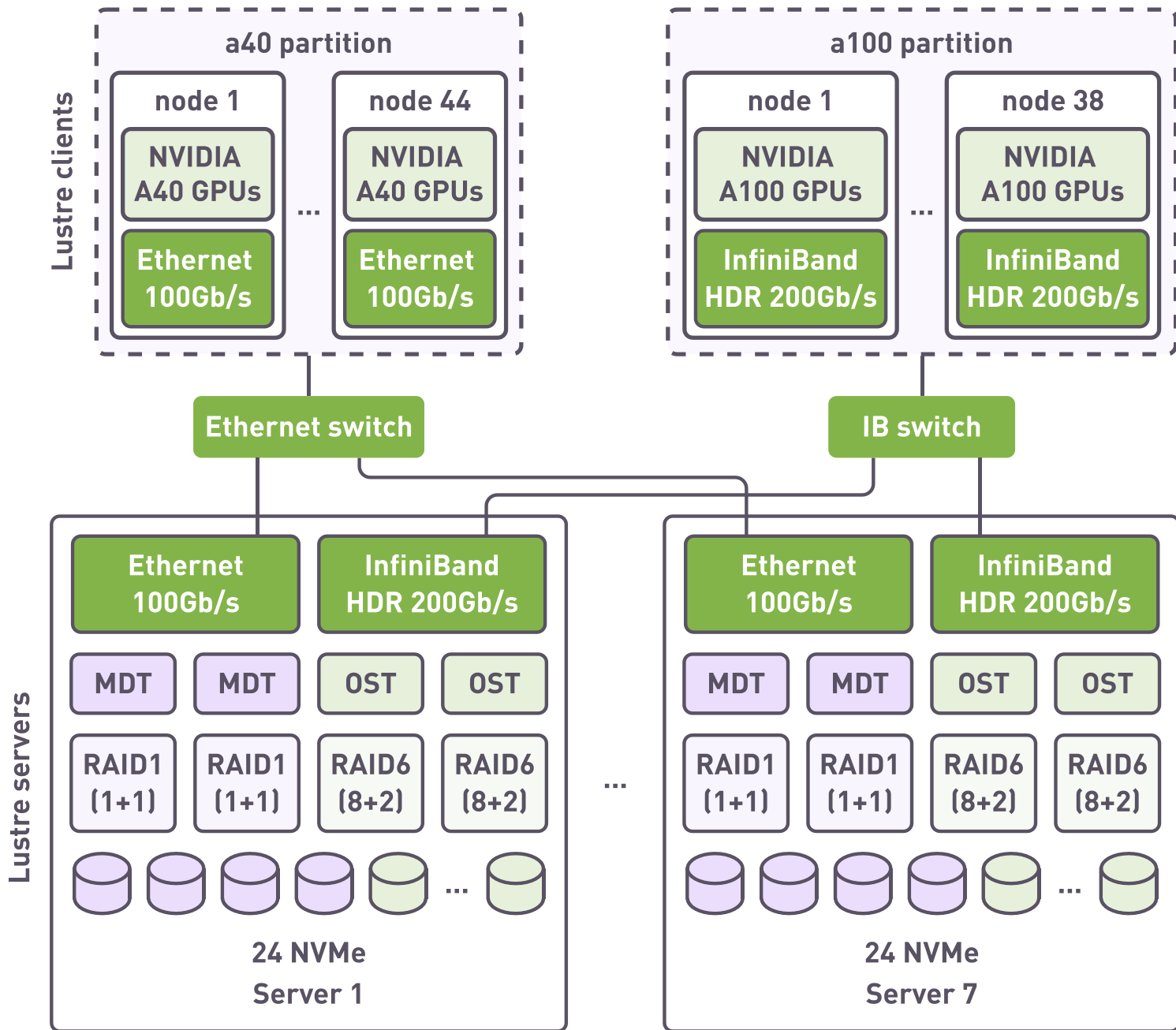




FAU – Hardware Storage Setup

7 servers:

- 2x AMD EPYC 7713 64-Core CPU
- 256 GB DDR4-3200
- 1x NVIDIA ConnectX-6 100Gbit/s Ethernet
- 1x NVIDIA ConnectX-6 200Gbit/s HDR InfiniBand
- 24x Intel D7-P5520 7.68TB PCIe4 NVMe
- Each server runs the following RAID config:
 - 2x MDT: xiRAID-1, ss=64k
 - 2x OST: xiRAID-6 (8+2), ss=64k
- Lustre 2.15.5
- AlmaLinux 8.10



FAU: Lustre Cluster Configuration

Lustre has 14 MDTs, 14 OSTs and ~775TB usable capacity

FAU: xiRAID +Lustre PoC Performance Test Results

- Original implementation: CephFS with 3-way replication
- 4 server nodes were migrated to xiRAID+Lustre and compared with CephFS over 3 servers

8 clients, 4 NVMe nodes with xiRAID6 + Lustre

(theoretical network limit via IB: $4 \times 25\text{GB/s} = 100\text{GB/s}$)

[RESULT] ior-easy-write **62.396781** GiB/s : time 480.873 seconds

[RESULT] ior-easy-read **84.449550** GiB/s : time 355.313 seconds

8 clients, 3 NVME nodes with CephFS

(theoretical network limit via Ethernet: $8 \times 3.125\text{GB/s} = 25\text{GB/s}$, clients only have 25Gbit/s Ethernet)

[RESULT] ior-easy-write **10.507753** GiB/s : time 311.646 seconds

[RESULT] ior-easy-read **21.787384** GiB/s : time 150.313 seconds

FAU: Pros of xiRAID + Lustre vs CephFS

Higher throughput

3-4x higher read-write throughput and metadata performance

More capacity

2x more usable capacity:

430TB → **≈775TB**
Ceph 3-way replica xiRAID + Lustre

InfiniBand RDMA support


FAU: what can be improved?

High Availability for server crash

- The system originally was not designed for Lustre, so it has no HA
- To implement HA, customer should use SBB systems or NVMe-oF EBOF

Disclaimer:
we will address this
challenge shortly

Partner's PoC for the
upcoming NDA project:
best performance with
Lustre + xiRAID

The background features a dark, textured surface with a glowing, grid-like pattern of white and yellow points. A bright, rainbow-colored light streak curves across the right side of the image, transitioning from red to orange, yellow, and green.

Cluster configuration

14 Servers

- AMD EPYC 9334 32-Core
- 384 GB DDR5-4800
- 2x NVIDIA ConnectX-7 NDR400
- 12x Micron 7450 Pro 3.84TB PCIe4 NVMe with specs:
 - Sequential:
 - 6800 MB/s - 128KB read
 - 4000 MB/s - 128KB write
 - Random:
 - 1000K IOPS - 4K read
 - 180K IOPS - 4K write

Each server runs the following RAID config:

- 1x MDT: xiRAID-1, ss=16k
- 1x OST: xiRAID-6 (8+2p), ss=64k

In total, Lustre has 14 MDTs, 14 OSTs and ~388TB usable capacity

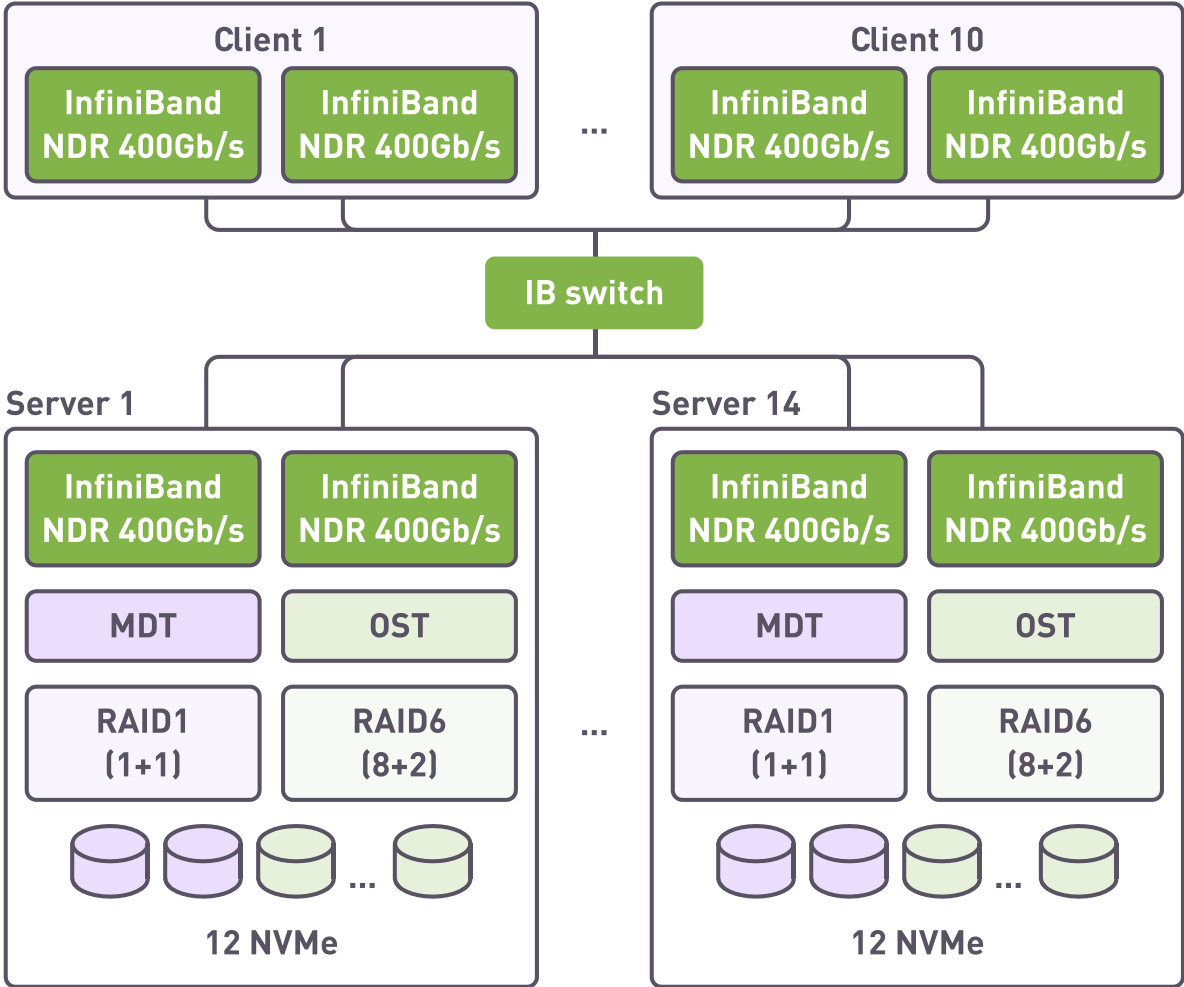
10 Clients:

- AMD EPYC 9334 32-Core
- 384 GB DDR5-4800
- 2x NVIDIA ConnectX-7 NDR400

in cooperation with



Cluster configuration



Theoretical performance

Read:

number of servers \times number of OST drives \times drive performance

$14 \times 10 \times 6700 \text{ MB/s} =$

916 GB/s

Write:

number of servers \times number of OST data drives \times drive performance

$14 \times 8 \times 3850 \text{ MB/s} =$

431 GB/s

IO500 results: 3rd fastest Lustre based system in the 10 Node research category

IO500 version io500-isc24_v3 (standard)

```
[RESULT]          ior-easy-write          328.653087 GiB/s : time 442.976 seconds
[RESULT]          mdtest-easy-write        658.996777 kIOPS : time 333.860 seconds
[          ]          timestamp            0.000000 kIOPS : time 0.000 seconds
[RESULT]          ior-hard-write           7.751412 GiB/s : time 442.857 seconds
[RESULT]          mdtest-hard-write        60.926374 kIOPS : time 310.050 seconds
[RESULT]          find                    18084.218841 kIOPS : time 13.167 seconds
[RESULT]          ior-easy-read            592.703319 GiB/s : time 245.650 seconds
[RESULT]          mdtest-easy-stat         2671.510444 kIOPS : time 83.093 seconds
[RESULT]          ior-hard-read            31.345065 GiB/s : time 109.503 seconds
[RESULT]          mdtest-hard-stat         2476.267432 kIOPS : time 8.599 seconds
[RESULT]          mdtest-easy-delete        458.735704 kIOPS : time 479.526 seconds
[RESULT]          mdtest-hard-read         671.807887 kIOPS : time 29.004 seconds
[RESULT]          mdtest-hard-delete        60.347770 kIOPS : time 312.913 seconds
[SCORE ] Bandwidth 82.943292 GiB/s : IOPS 739.394914 kiops :
```

TOTAL 247.644601

Elbencho results

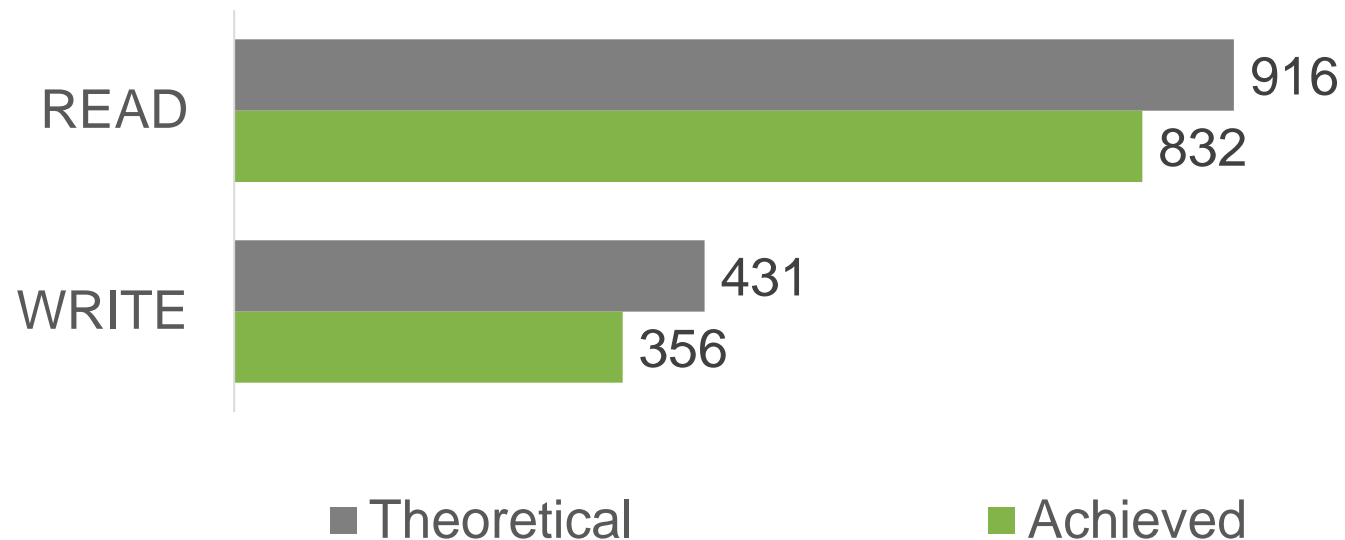
```
# ./bin/elbencho --hosts $(cat hosts) -t 28 -w -r -b 64M -s 200g --direct --  
iodepth 64 /mnt/elbencho/testfile{01..280}
```

OPERATION	RESULT TYPE	FIRST DONE	LAST DONE
=====	=====	=====	=====
WRITE	Elapsed time	: 2m36.535s	2m46.715s
	IOPS	: 5443	5374
	Throughput MiB/s	: 348414	343963
	Total MiB	: 54539264	57344000

READ	Elapsed time	: 58.640s	1m45.093s
	IOPS	: 12699	8525
	Throughput MiB/s	: 812748	545648
	Total MiB	: 47660096	57344000

Elbencho results: comparison with theoretical maximum

Performance Test Results, GB/s:



Read Efficiency: **90,8%**

Write Efficiency: **82,6%**

High Availability in xiRAID Installations for Lustre

The background of the slide is a dark, abstract composition. On the right side, there is a glowing, grid-like structure that resembles a molecular lattice or a data network. This structure is illuminated with a gradient of colors, starting with a bright red and orange glow that transitions into a vibrant green towards the right edge. The overall effect is one of high-tech, digital energy.

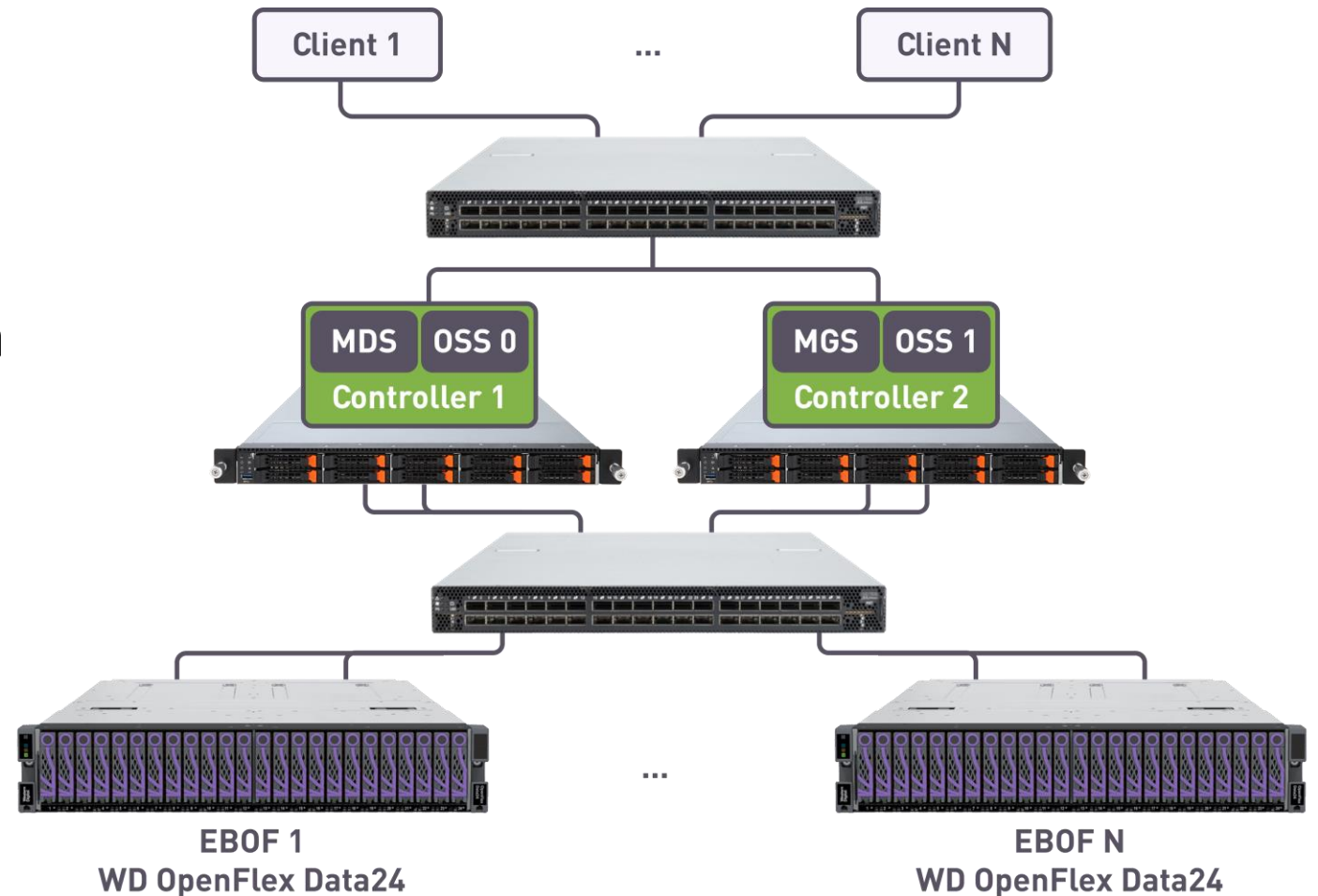
xiRAID High Availability Architecture

- Lustre HA approach is based on Pacemaker clustering
- xiRAID Classic 4.1 includes Pacemaker agent for an HA cluster integration
- Shared RAID drives required
- Dual-node clusters are supported
- Multi-node cluster support planned
- Target system architectures:
 - SBB systems with dual-ported drives
 - EBOF connected to the cluster nodes

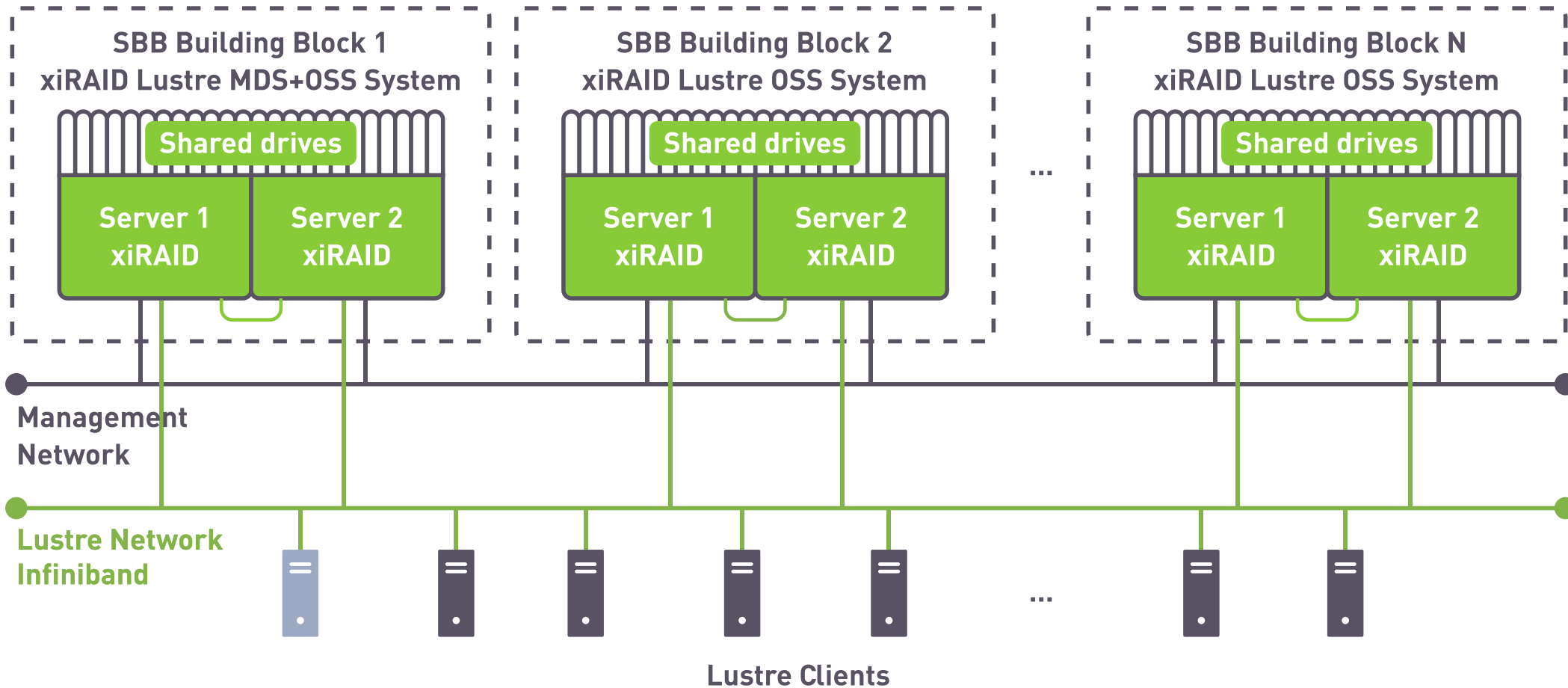


Disaggregated Lustre Solution with xiRAID Classic on WD Data24

- A high-performance, scalable storage cluster
- Offers large-scale storage with robust failover capabilities and high throughput
- A solution brief is available



xiRAID High Availability Architecture: SBBs as Lustre Building Blocks





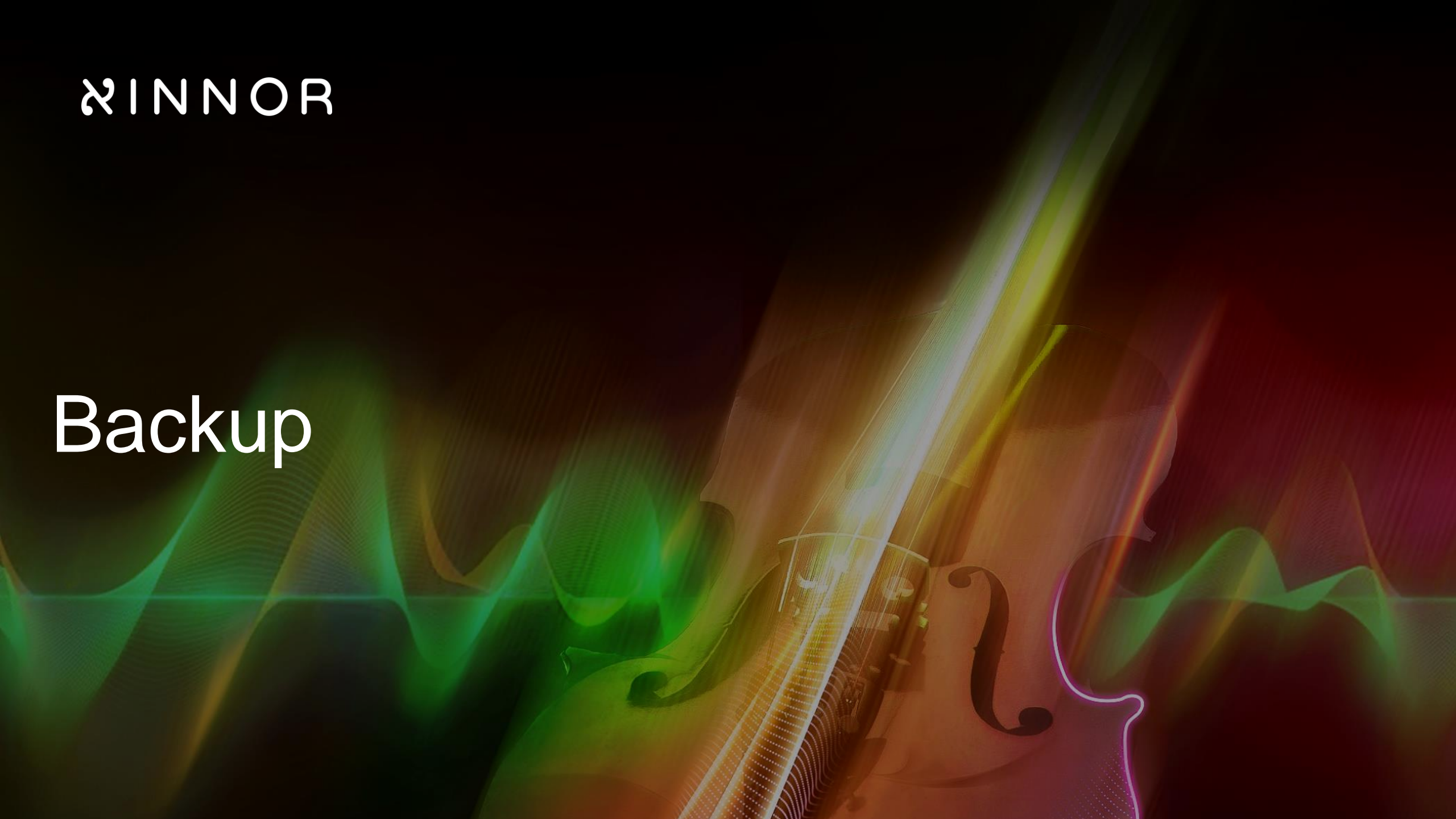
Our article: Building a High-Performance, Highly Available Lustre Solution with xiRAID Classic 4.1 on a Dual-Node System with Shared NVMe Drives

Prove it yourself:
<https://xinnor.io/>



INNOR

Backup



XINNOR

Xinnor Lustre Solution for Cloud Environments

A violin is the central focus of the image, rendered in a semi-transparent, glowing style. The body of the violin is a warm, golden-brown color, while the f-hole and scroll are dark. A vibrant, multi-colored aura surrounds the instrument, with green and purple light waves emanating from it. The background is a dark gradient with soft, glowing light trails in green and purple, creating a sense of motion and energy. The overall aesthetic is modern and artistic, suggesting a fusion of traditional craftsmanship with cutting-edge technology.

xiRAID Opus (Optimized Performance in User Space)

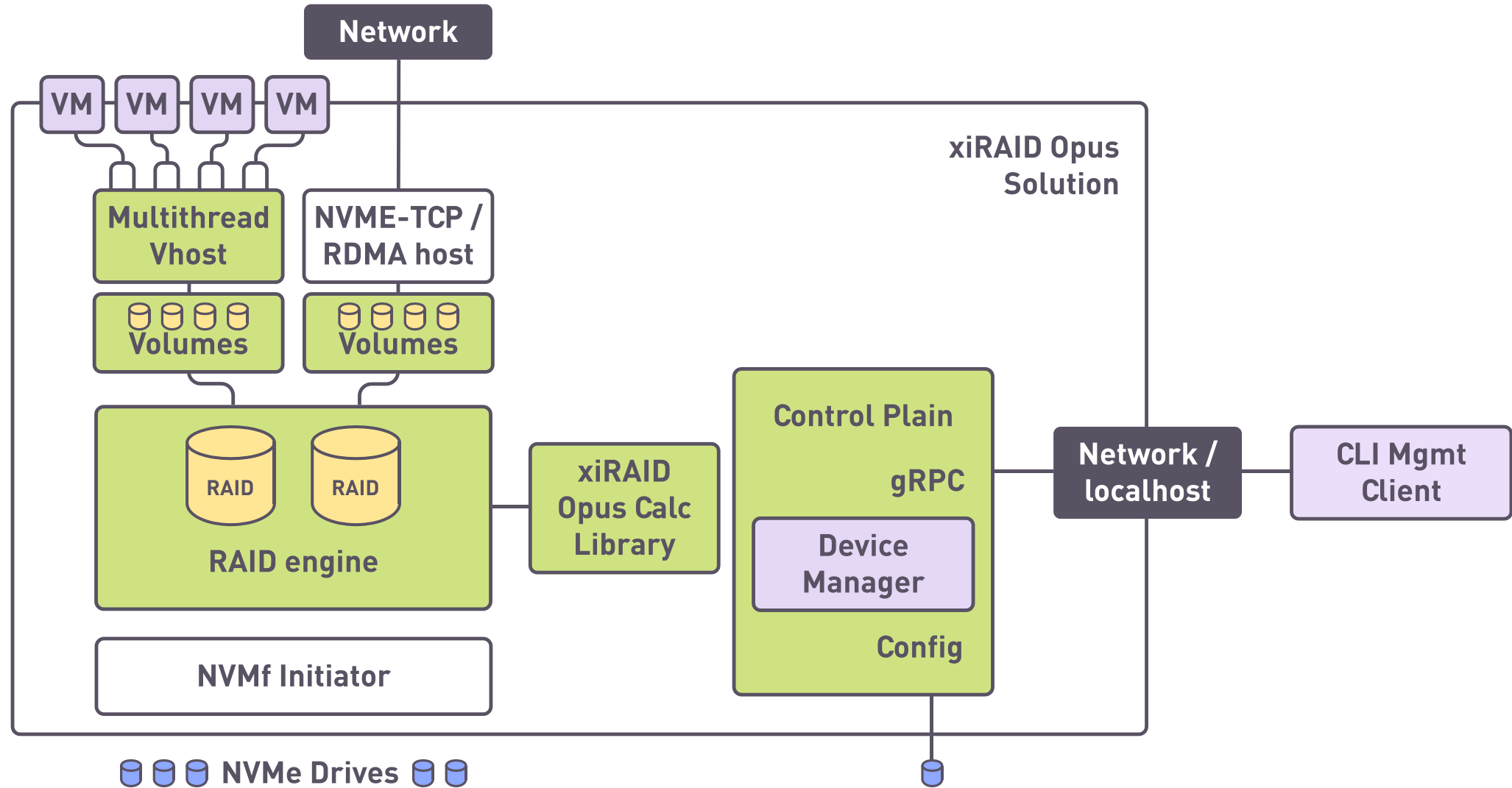
Engine designed for high-performance storage data paths in virtualized environments.

1. Creation of RAID-protected volumes.
2. Provisioning of volumes to VMs.
3. Performance Enhancements:
 - Polling: Reduces latency by actively checking for I/O completions.
 - Zero-Copy: Eliminates unnecessary data copying, increasing throughput.

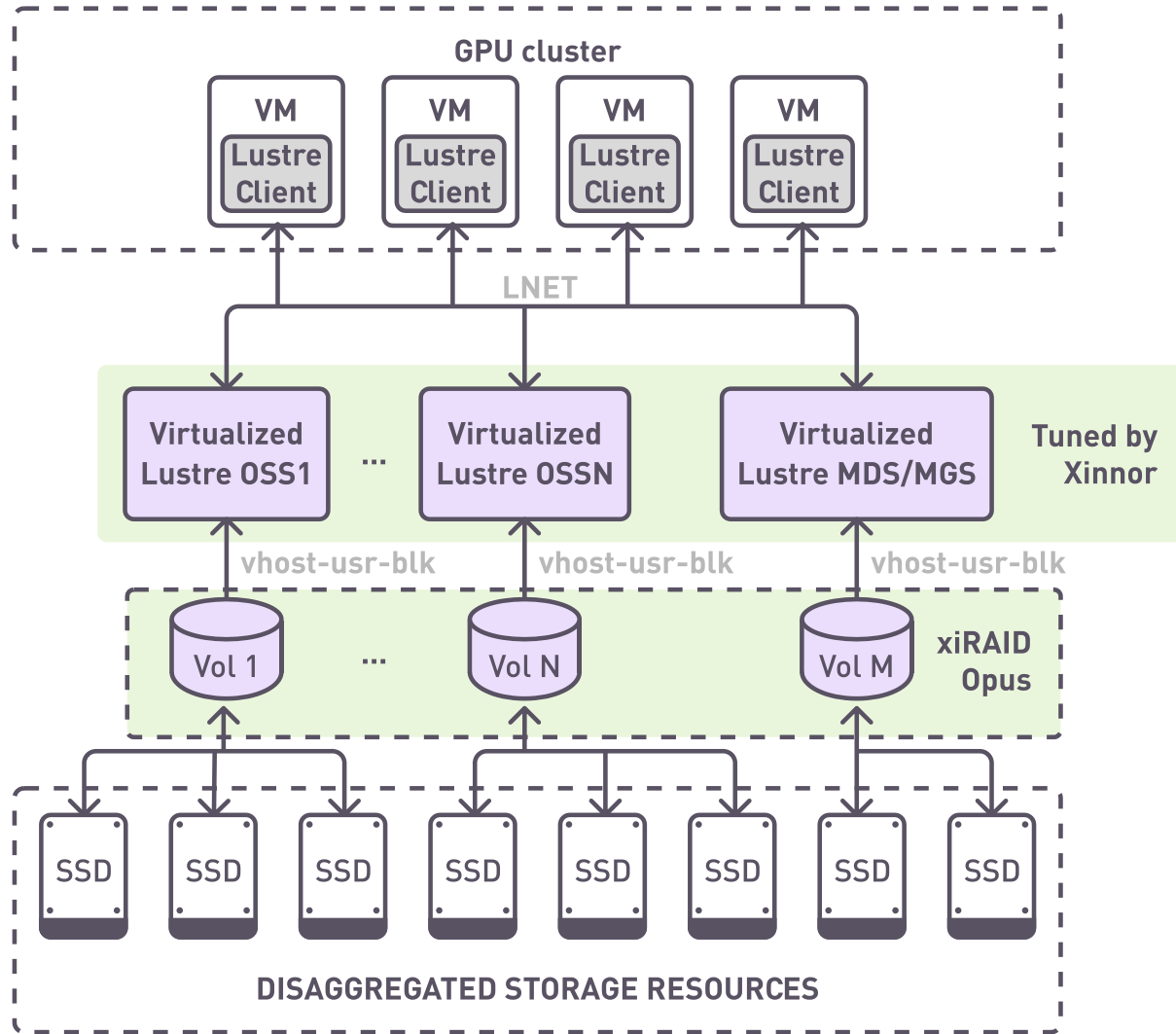
	Measured single drive performance	2x RAID5 theoretical performance	xiRAID 2x RAID5 performance	Efficiency
4K Random Read (M IOPS)	2,7	65	65	100%
4K Random Write (M IOPS)	0,7	8	8	100%
Sequential Read (GB/s)	14	336	310	92%
Sequential Write (GB/s)	6,75	149	144	97%

xiRAID demonstrates world-record performance with 24 Kioxia CM7 PCIe 5 NVMe SSD drives

xiRAID Opus Architecture

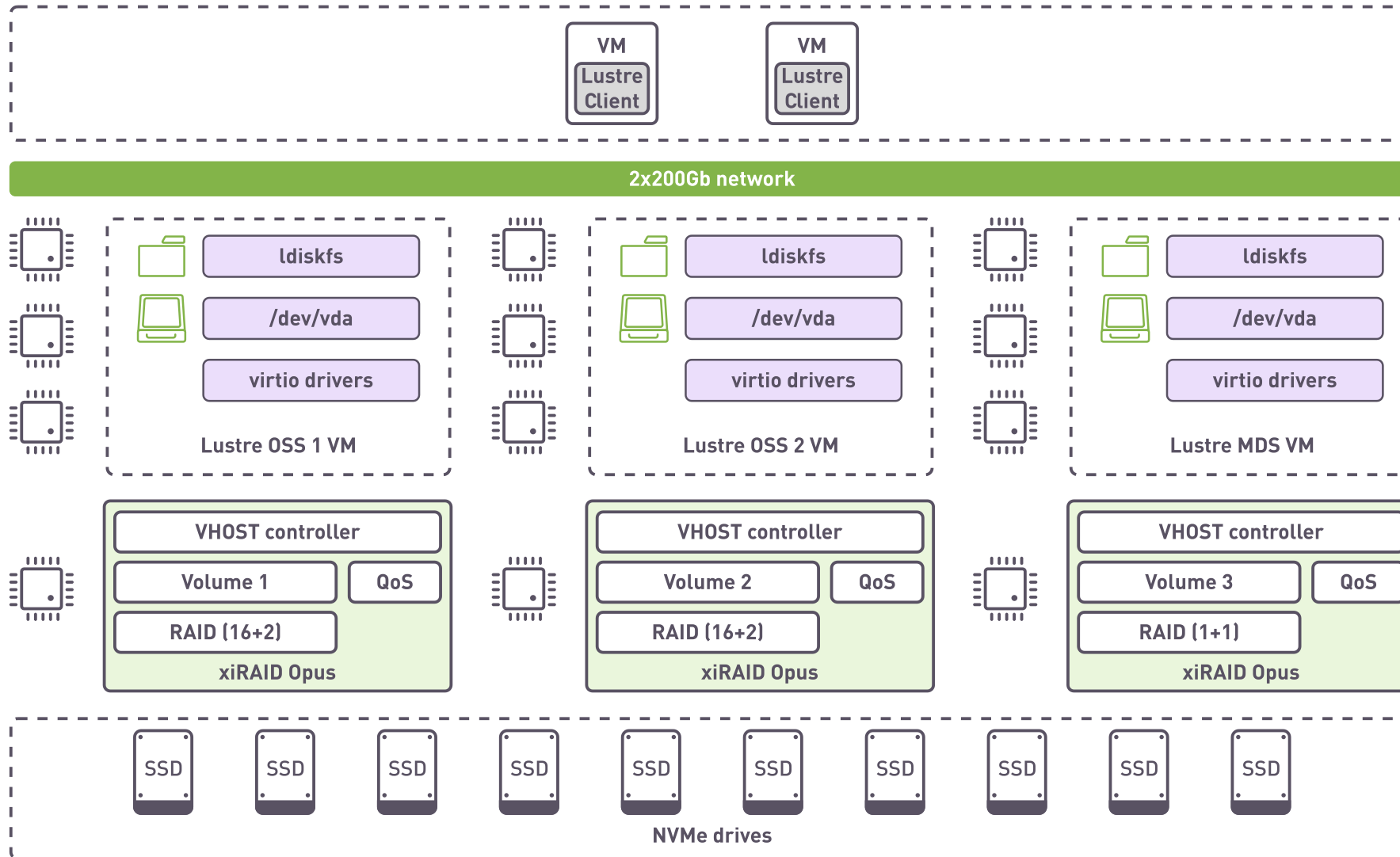


Lustre in Cloud Environments



- Lustre is a well-known FS that is primarily used for HPC workloads
- Provides good scalability and performance for data intensive workloads
- Provides HA over shared storage

Testing Environment Details



CPU: 64-Core Processor per node (AMD 7702P)

Memory: 256 GB RAM per Node

Networking: 1 x MT28908 Family [ConnectX-6] per node

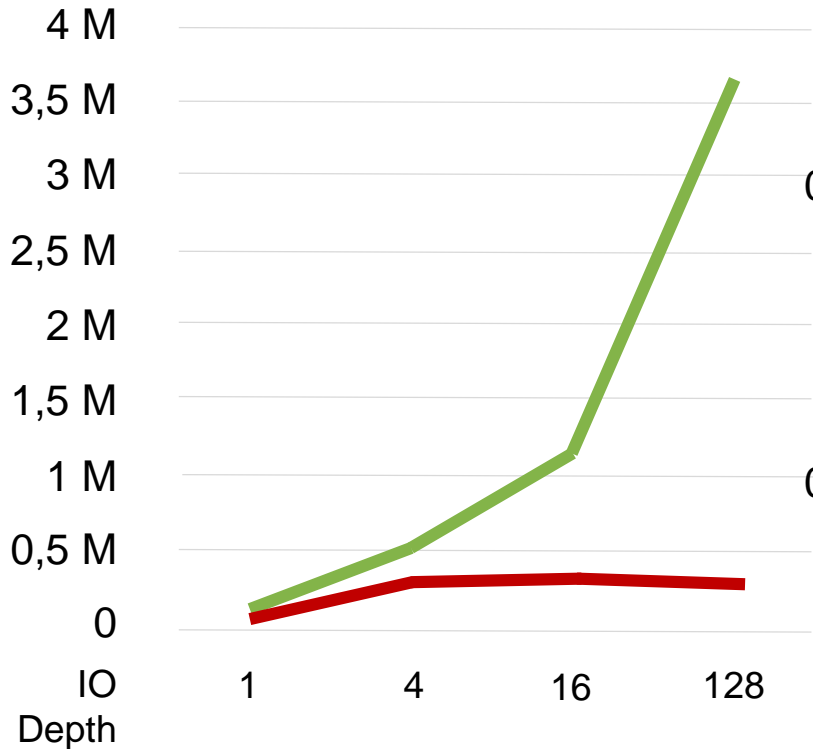
Drives: 24x KIOXIA CM6-R 3.84TB (Gen 4)

Aggregated drive performance for each node:

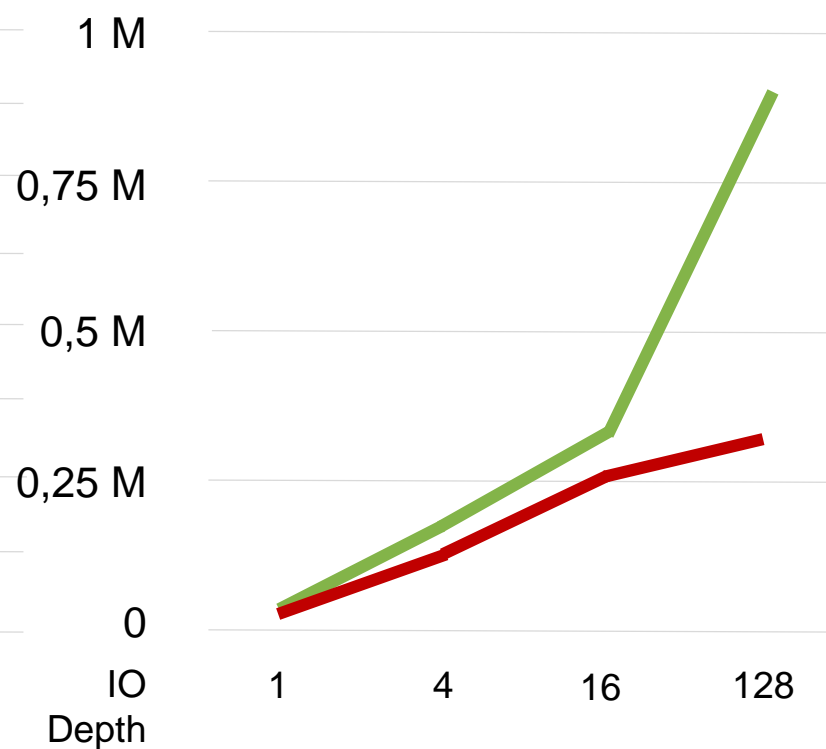
- 9M IOps 4k RR
- 3M IOps 4k RW
- 70 GBps 128k SW, SR

Lustre Solution Performance

Random read, 32 jobs, IOps



Random write, 32 jobs, IOps




Sequential read 1M, 32 jobs:


- with xiRAID Opus: **44 GB/s**

Sequential write 1M, 32 jobs:

- with xiRAID Opus: **43 GB/s**

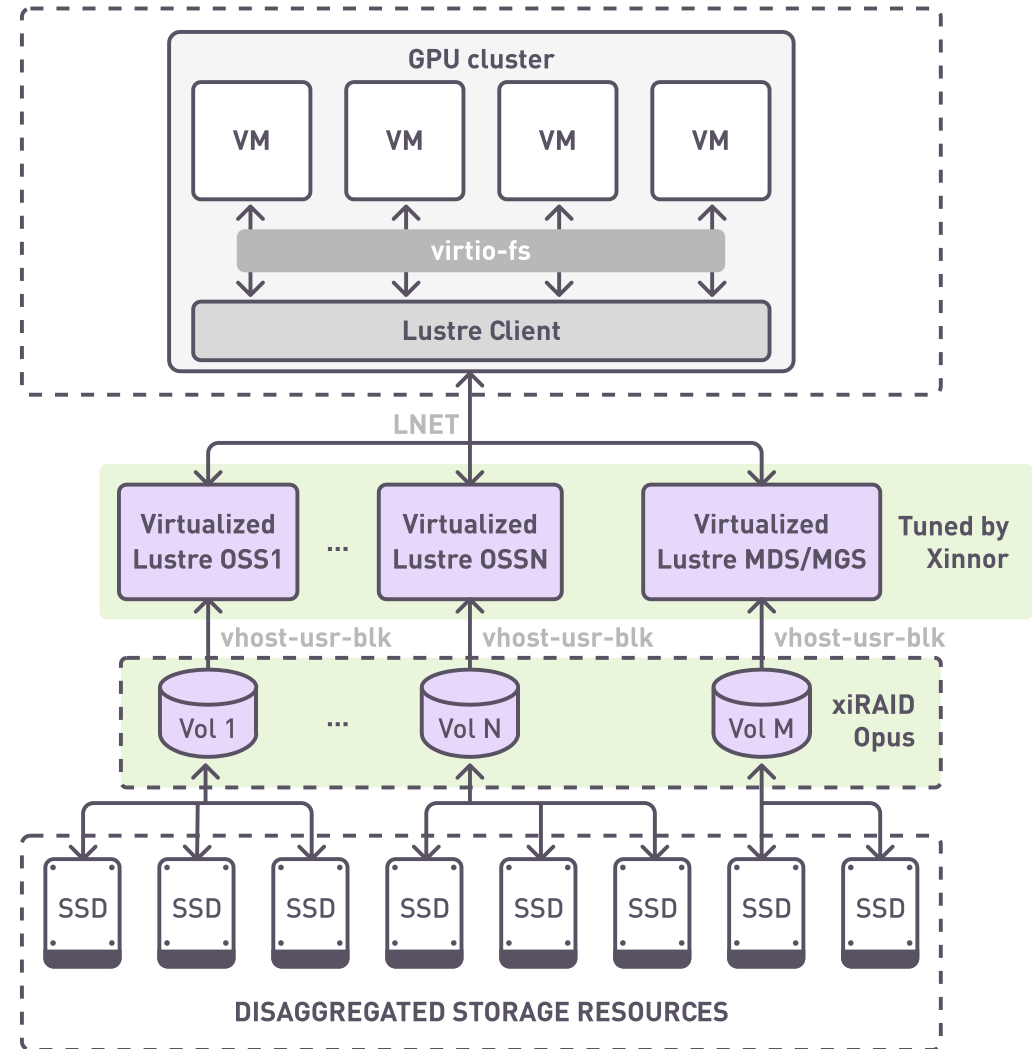
These results can be achieved with multithreaded vhost-user-blk only!

 Lustre w/ xiRAID Opus (RAID6)

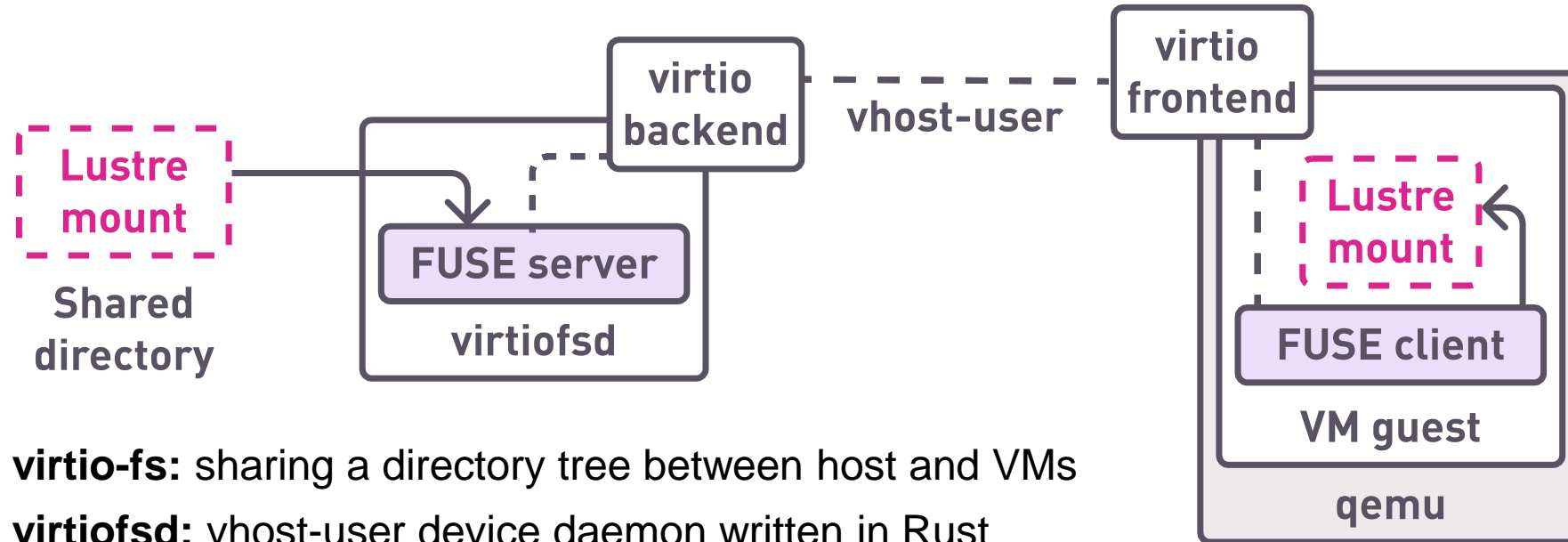
 Lustre w/o xiRAID Opus (RAID0)

Reducing complexity of Lustre administration

- Virtiofs allows to share mounted FS on the host with VMs
- No client software or specialized networking configuration needed



Virtio FS Upsides = simplicity

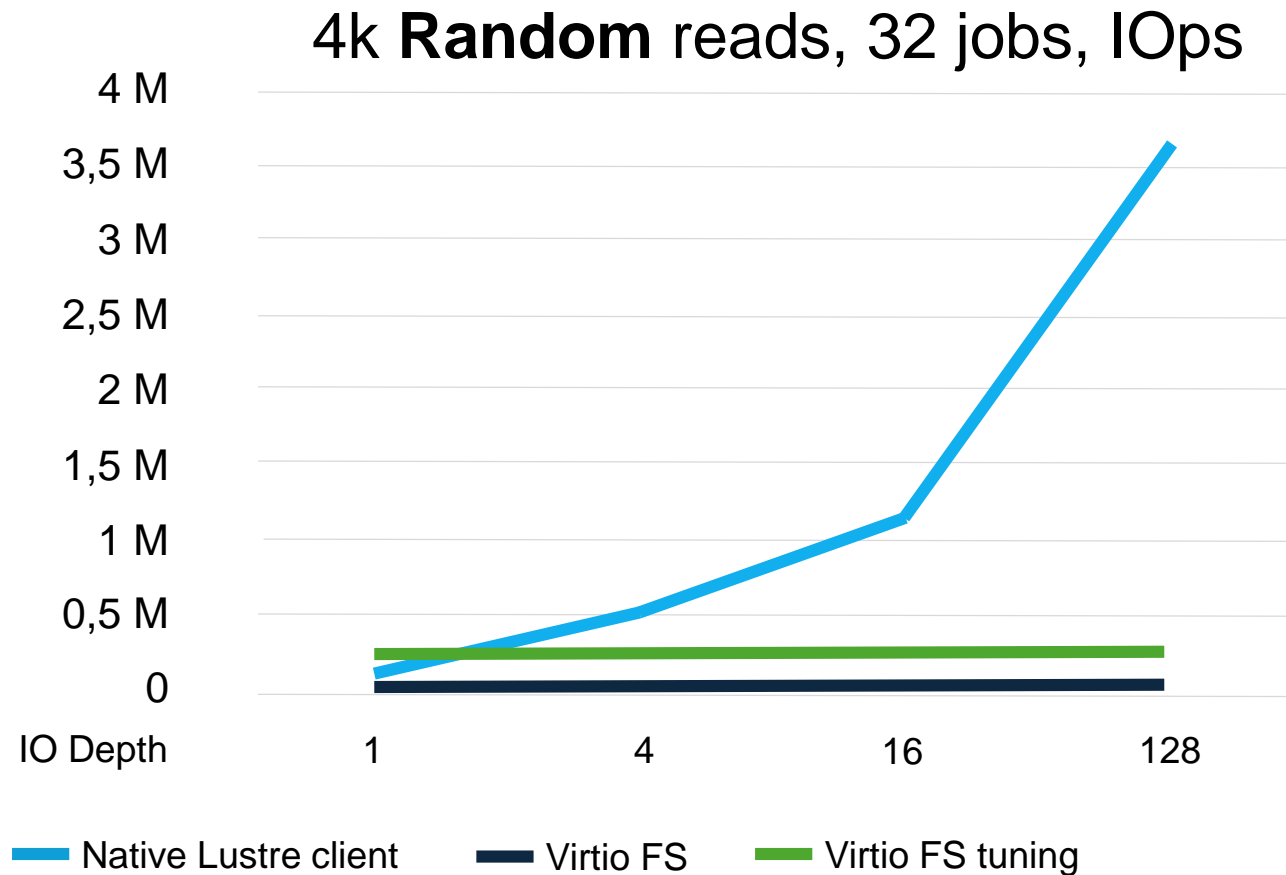


virtio-fs: sharing a directory tree between host and VMs

virtiofsd: vhost-user device daemon written in Rust

We can hide all the complexity of parallel file system setup from the client behind VIRTIOFS!

Tuned Virtio FS = performance results



Sequential read 1M, 32 jobs:

- Native Lustre client: 44 GB/s
- Virtio FS: 9 GB/s
- Tuned Virtio FS: 44 GB/s

Sequential write 1M, 32 jobs:

- Native Lustre client: 43 GB/s
- Virtio FS: 7 GB/s
- Tuned Virtio FS: 44 GB/s

Outcomes = Xinnor Lustre Solution can perform in Cloud

1. Performance:

- Even with only two virtualized OSS Lustre delivers strong results for sequential and random I/O operations (AIO).
- It is essential to have high-performance block devices passed through to the OSS and MDS virtual machines.
- **xiRAID Opus solves this challenge.**

2. Skill requirements:

- Requires a high level of expertise to configure the system and client VMs.
- VirtioFS can reduce complexity:
 - For use-cases that require only sequential workload patterns.

3. Xinnor can deliver Lustre solution for Cloud Environments.

Reach out to Xinnor for a POC.