



XINNOR

IBM Scale acceleration

IBM Scale User Group London 2026

Dmitrii Livshits – CEO, livshits@xinnor.io

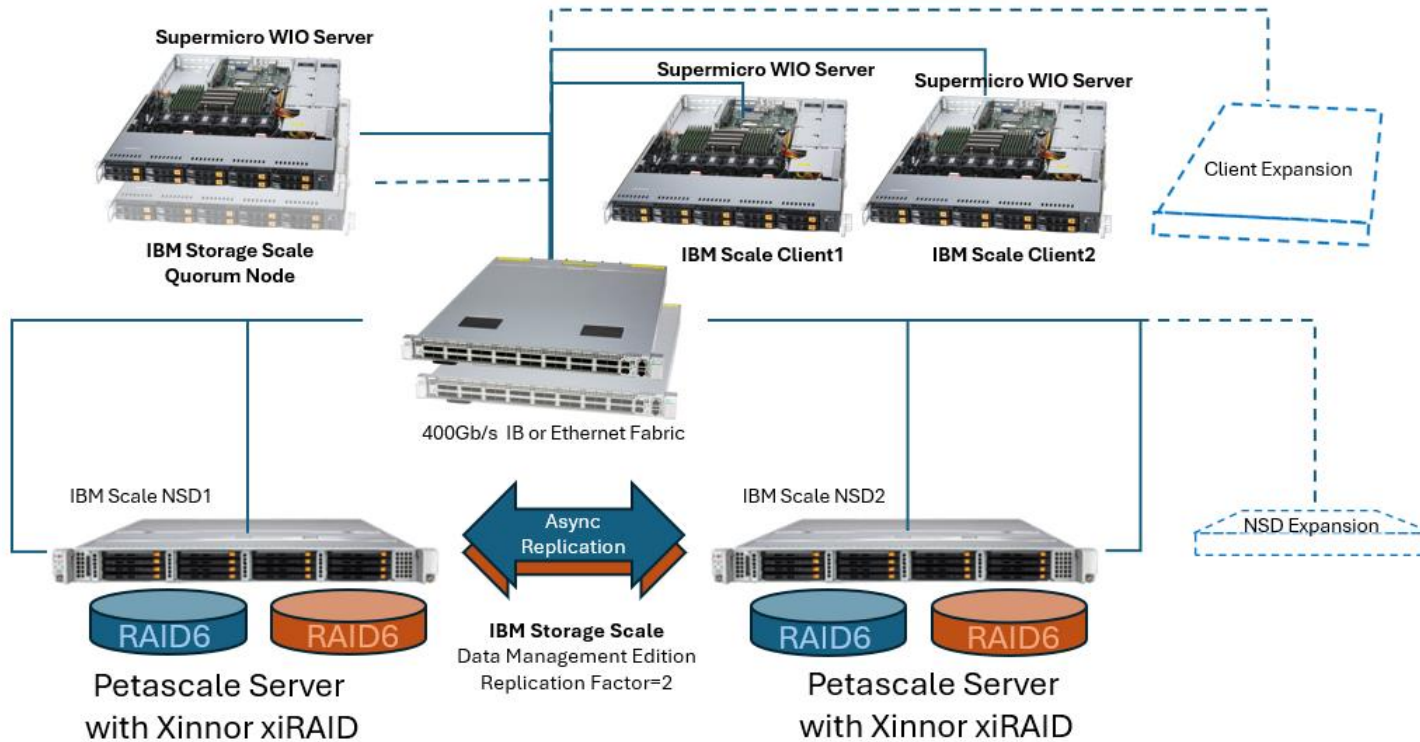
Daniel Landau – Senior Solution Architect, landau.d@xinnor.io

IBM Scale DME node protection and acceleration



Reference Architecture

Reference Architecture



Data Nodes: 2x Supermicro SSG-122B-NE316R

- CPU: Intel Xeon with min 32 cores
- Memory: 256GB DRAM
- Storage: 6x PCIe Gen5 SSDs
- Networking: 2x Mellanox ConnectX-7

Control and Client Nodes:

- AS-1114S-WN10RT-1 (Quorum)
- AS-1114S-WN10RT-2/-3 (Client Nodes)
 - Networking: 1x Mellanox ConnectX-7

System Software: RHEL 9

Drive-Level Protection: xiRAID Classic 4.3 in RAID6

Multinode-Level Protection: IBM Storage Scale (GPFS) based on 5.2.0 or later; NSD (Network Shared Disks) Replication Factor 2

Real-life use case

Storage requirements:

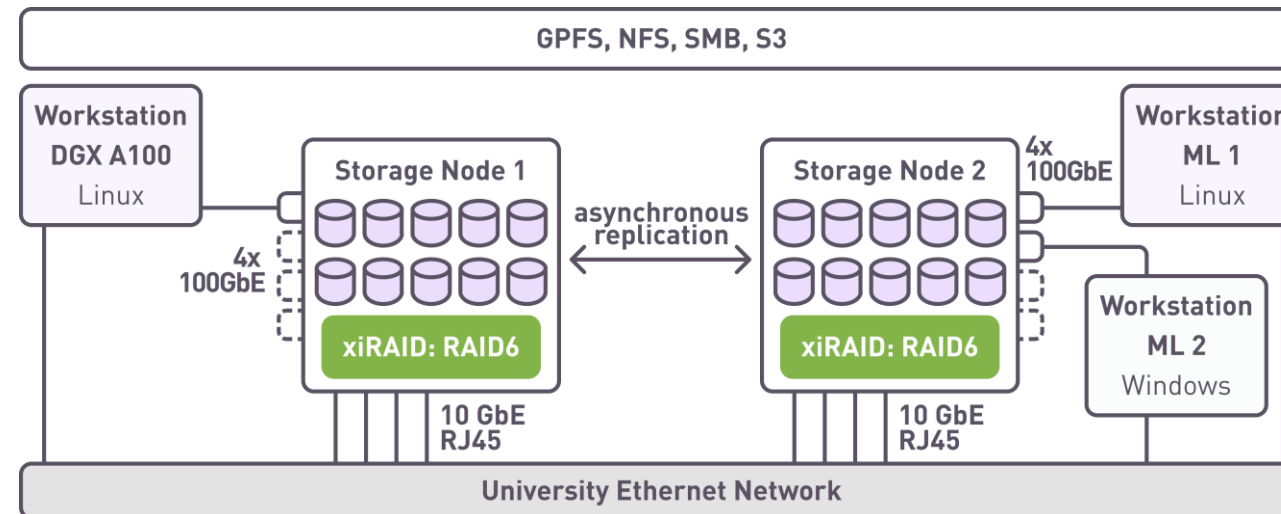
- High-Performance:
 - to saturate the network bandwidth to the client (Nvidia DGX and ML workstations)
- Data Availability and Integrity:
 - capable of surviving hardware failures
- Geographical Distribution:
 - to synchronize the data between the 2 sites of the campus for DR
- Scalability:
 - to accommodate future growth in data and compute resources.
- Cost efficiency



Aalen University

High-performance storage for ML based on

- IBM Scale Data Management Edition
- xiRAID RAID6 over 2 servers with 10 drives each
- Standard Supermicro NVMe server



<https://xinnor.io/case-studies/aalen/>

Single Node Performance: 10 drives in RAID6

	Sequential Read	Sequential Write
Raw drive performance	92 GB/s	55 GB/s
Storage Scale + xiRAID	80.2 GB/s	42.8 GB/s
Overall efficiency	87 %	78 %*

* All 10 drives are considered when calculating the efficiency

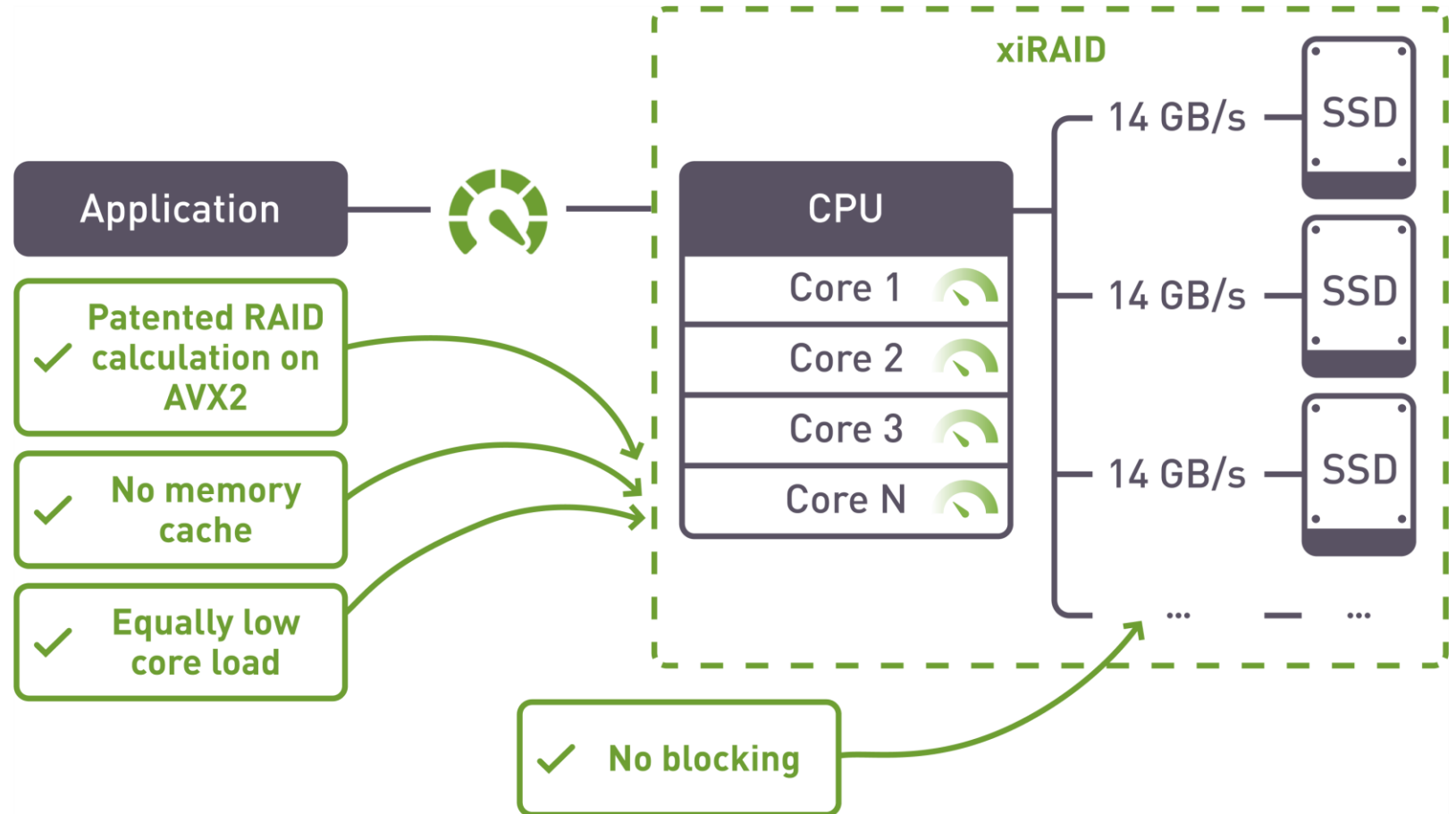
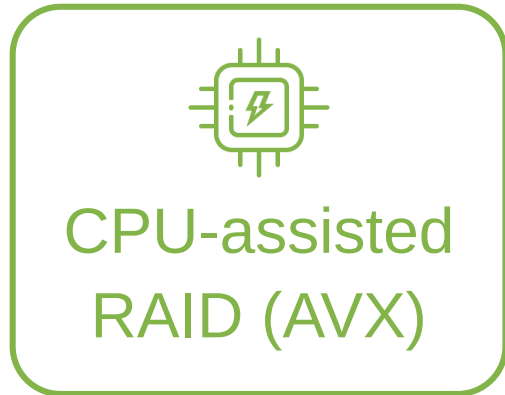
- xiRAID delivers throughput speeds very close to the raw performance of the drives.
- The solution saturates the network.

GPFS FS with gpfsperf:

```
```\n/usr/lpp/mmfs/samples/perf/gpfsperf write seq 120Gfile -n 120g -r 4m\n-th 10 -fsync -dio\n  recSize 4M nBytes 120G fileSize 120G\n  nProcesses 1 nThreadsPerProcess 10\n  file cache flushed before test\n  using direct I/O\n  offsets accessed will cycle through the same file segment\n  not using shared memory buffer\n  not releasing byte-range token after open\n  fsync at end of test\n  Data rate was 42836998.24 Kbytes/sec, Op Rate was 10213.14\n  Ops/sec, Avg Latency was 0.977 milliseconds, thread utilization\n  0.997, bytesTransferred 128849018880
```

```
\n/usr/lpp/mmfs/samples/perf/gpfsperf read seq 120Gfile -n 120g -r 4m\n-th 10 -fsync -dio\n recSize 4M nBytes 120G fileSize 120G\n nProcesses 1 nThreadsPerProcess 10\n file cache flushed before test\n using direct I/O\n offsets accessed will cycle through the same file segment\n not using shared memory buffer\n not releasing byte-range token after open\n Data rate was 80189171.64 Kbytes/sec, Op Rate was 19118.59\n Ops/sec, Avg Latency was 0.513 milliseconds, thread utilization\n 0.980, bytesTransferred 128849018880
```

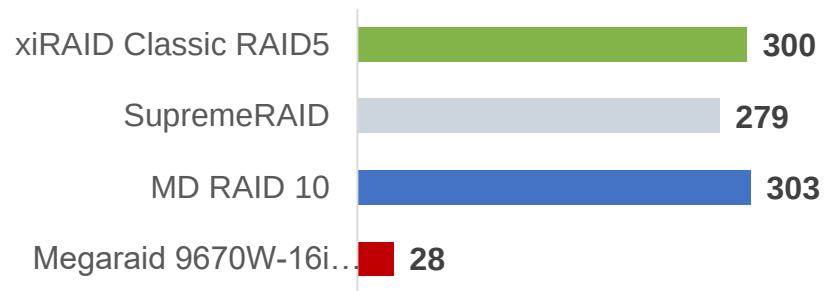
# xiRAID Classic



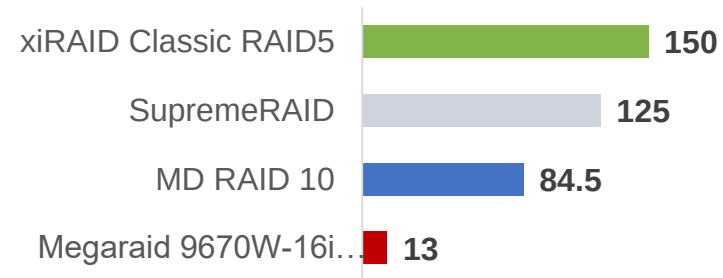
# Benchmark

xiRAID, SupremeRAID and MDRAID over 24 drives while MegaRAID over 16 drives

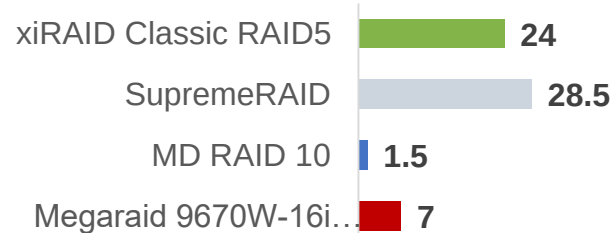
## Sequential read, GB/S



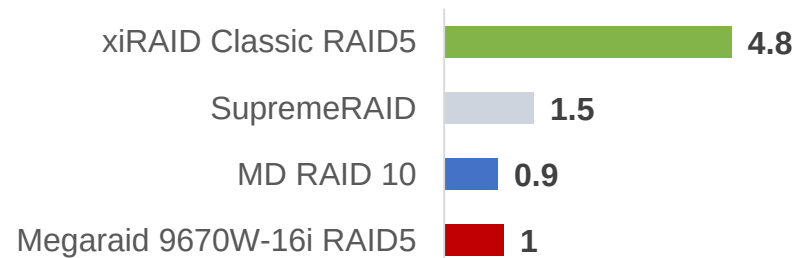
## Sequential write, GB/S



## Random read, M IOPS



## Random write, M IOPS



## Sources:

<https://xinnor.io/blog/pcie-gen-5-storage-array-optimization-overcoming-the-obstacles-to-maximize-performance/>

[www.supermicro.com/solutions/Solution\\_Brief\\_GRAID\\_SupremeRAID\\_RAID6.pdf](http://www.supermicro.com/solutions/Solution_Brief_GRAID_SupremeRAID_RAID6.pdf)

<https://www.storagereview.com/review/broadcom-megaraid-9670w-16i-raid-card-review>

# QLC – Rebuild Without Workload

Rebuilding 1x Solidigm D5-P5336 61.44TB QLC

RAID engine	Rebuild time	Rebuild speed	WAF (lower is better)
mdraid	53h 40m	322MB/s	1.2
xiRAID Classic 4.3	5h 22m 10x faster rebuild	3.18GB/s 10x higher throughput	1.02 15% lower WAF

<https://www.solidigm.com/products/technology/raid-rebuild-with-xiraid-and-qlc-ssds.html>

# QLC – Rebuild With Workload

Rebuilding 1x Solidigm D5-P5336 61.44TB QLC in RAID 5 over 9 drives

RAID Engine	Rebuild time	Rebuild speed	WAF (lower is better)	Workload speed under rebuild
mdraid	>67 days	10.5 MB/s	1.58	Read: ~100MB/s Write: ~45MB/s
xiRAID Classic 4.3	53h 53m 25x faster rebuild	316 MB/s 30x higher throughput	1.21 23% lower WAF	Read: 44GB/s Write: 13GB/s 290-440x higher

<https://www.solidigm.com/products/technology/raid-rebuild-with-xiraid-and-qlc-ssds.html>

# NEW: BOM from IBM&Supernova&Xinnor

## Usable capacity:

- 2 storage node
  - 122TB or 245 TB
- 4 storage nodes:
  - 490 TB
- 6 storage nodes:
  - 737 TB

## What customer gets:

- Fully preconfigured solution ready to use immediately after shipment
- Single point of contact for support (no finger pointing)
- Unbeatable performance per TB
- Functional-rich enterprise parallel file system for small and medium-sized deployment
- Network-saturated storage for GPU-intensive compute

# Who can benefit from this solution

1. Customers looking for a high performance (all-flash) storage based on the most reliable POSIX file system with budget constraints.
2. Customers who want to repurpose their off-the-shelves NVMe servers getting enterprise-grade solution without investing in new HW

# Network (NVMeOF) block device for IBM Storage Scale acceleration

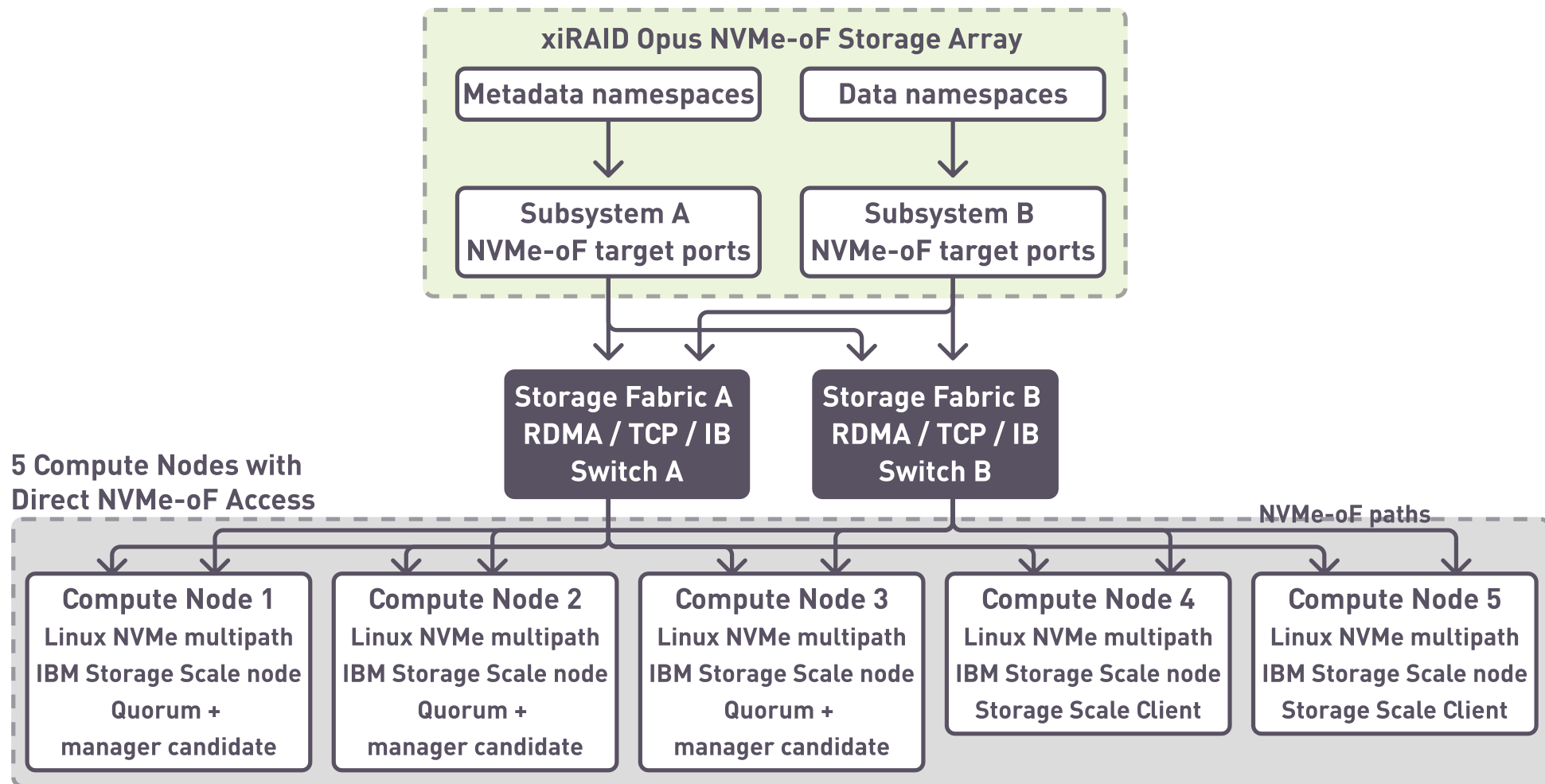


# Use cases

- **NVMeOF block device for ephemeral storage**
  - exposes remote NVMe capacity as protected block devices over NVMe-oF/RDMA for IBM Storage Scale.
  - eliminates the additional horizontal NSD data traffic.
- **NVMeOF block device for “hot” data**
  - xiRAID Opus adds a separate high-performance NVMe-oF/RDMA flash layer without replacing the existing Storage Scale infrastructure.
- **High-performance backend for DAT**
  - xiRAID Opus provides an alternative disaggregated NVMe-oF/RDMA flash layer that can deliver high-performance block volumes to IBM Storage Scale without requiring client-local NVMe or a dedicated IBM all-flash appliance.

# Use case 1. Storage Scale over External NVMe-oF

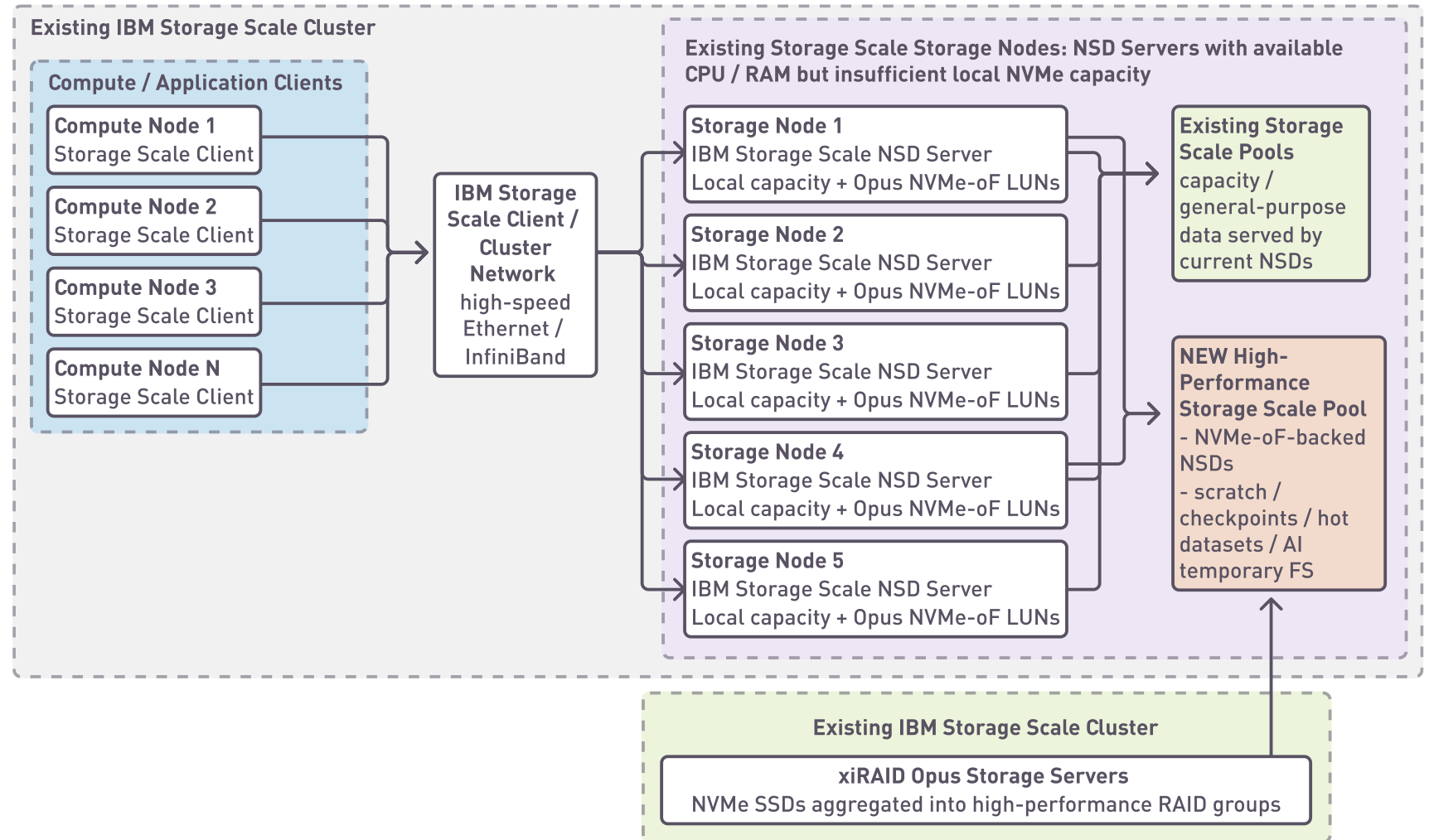
## Shared-Disk / No Dedicated NSD Servers



# Use case 2. IBM Storage Scale NSD Cluster

**Key idea:** reuse existing Storage Scale NSD servers as the file-system serving layer. Add external NVMe capacity through xiRAID Opus and expose it as new NVMe-oF block devices, then create new NSDs and place them into a dedicated high-performance Storage Scale pool.

**Result:** no new IBM Storage Scale storage servers are required. Existing NSD servers provide CPU, network and Scale services; xiRAID Opus provides additional protected NVMe performance/capacity.



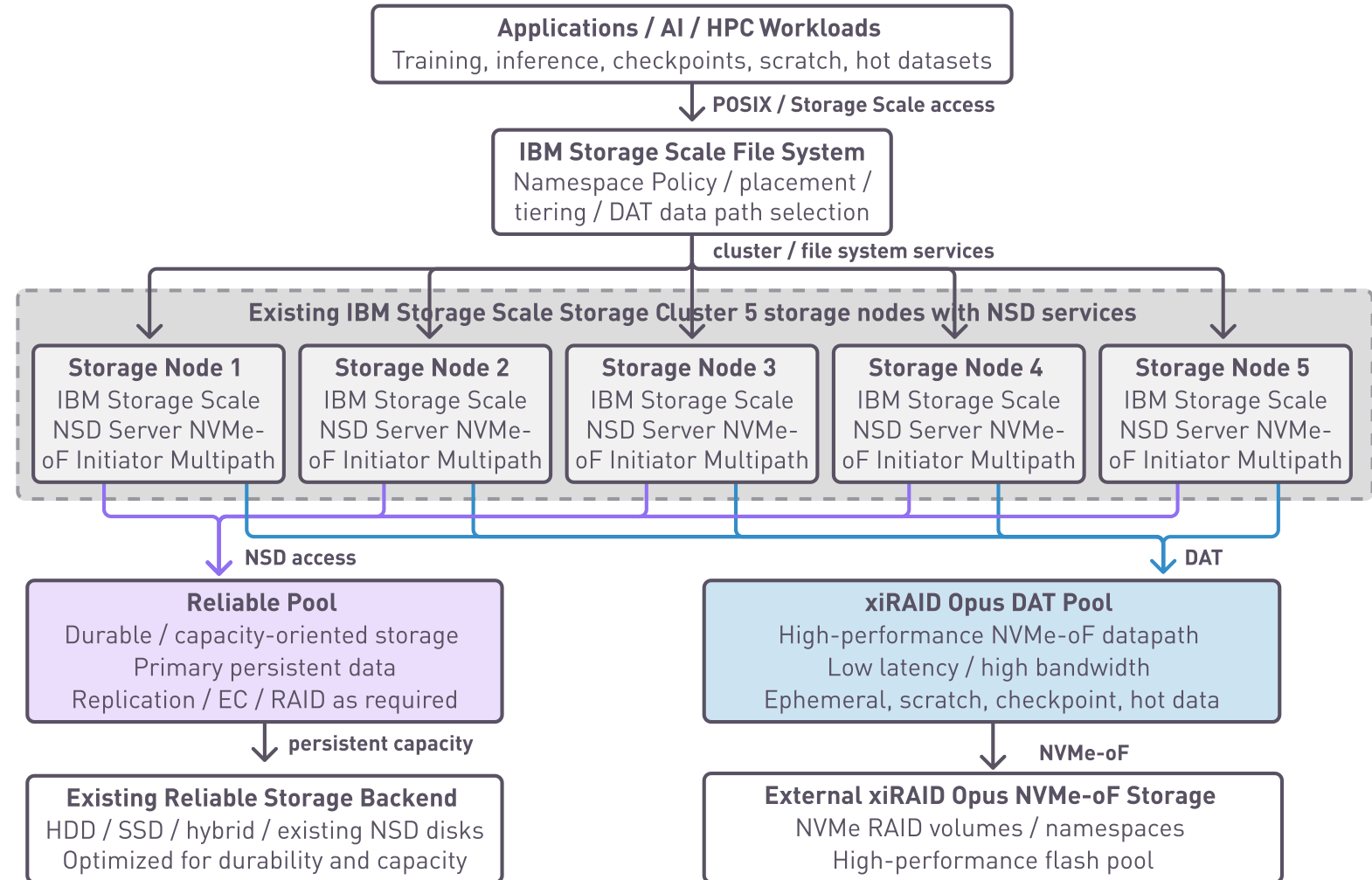
# Use case 3. Storage Scale with Reliable Pool and xiRAID Opus DAT Pool

## Design intent:

- Keep the existing 5 IBM Storage Scale storage nodes.
- Add high-performance capacity without buying new Scale servers.
- Reliable Pool remains the durable / persistent layer.
- xiRAID Opus pool is used as the high-performance DAT datapath.

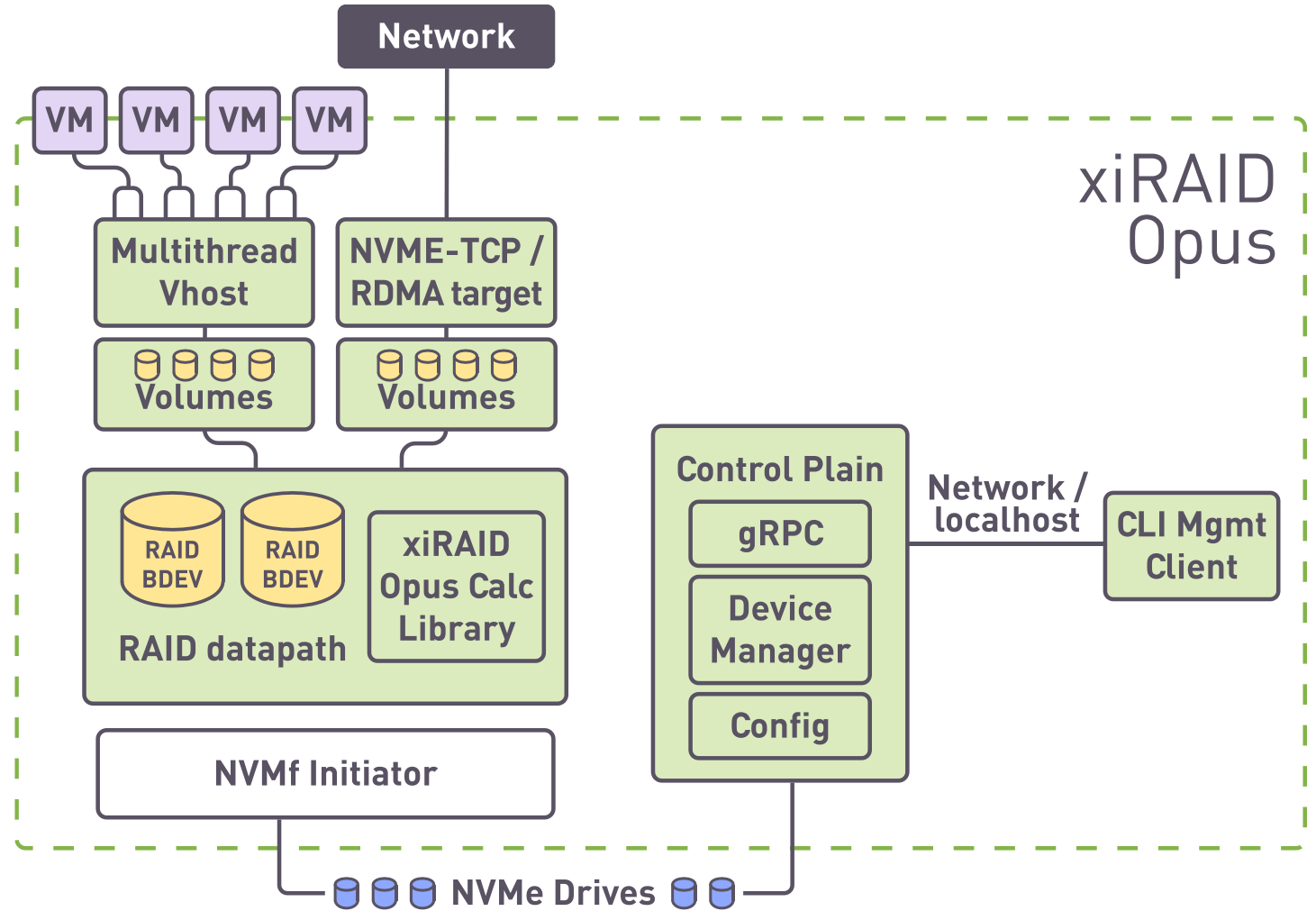
## Operational requirement:

- Opus namespaces must be exposed to all selected NSD servers.
- Use redundant fabrics and multipath.
- IBM Storage Scale consumes stable multipath devices as NSDs.
- Storage policies place hot / ephemeral data into the DAT pool.



# xiRAID Opus architecture

- xiRAID Opus is based on SPDK libraries to operate in **User Space** (data path and drive connection)
- xiRAID Opus uses **polled-mode** instead of interrupts (Kernel mode) to check when data is ready, reducing CPU cycles and maximizing performance
- CPU utilization: xiRAID Opus fully occupies **one or more CPU cores** to achieve max performance
- Unique **vectorized RAID Checksum calculation**:
  - AVX2 on x86
  - Vector registers on ARM



# xiRAID Opus performance from a single CPU core

Tested RAID6:

6+2 PCIe Gen5 drives

**Learn more:**

[Blog post: High performance with minimal resource consumption with the xiRAID Opus and Sandisk PCIe Gen5 drives](#)

**Sequential read:**

**100 GB/s**

**Sequential write:**

**16 GB/s**

**Random read:**

**1 M IOPS**

**Random write:**

**0.3 M  
IOPS**

*“Xinnor’s architecture demonstrates how innovative software design can unlock E-core capabilities for demanding storage applications.*

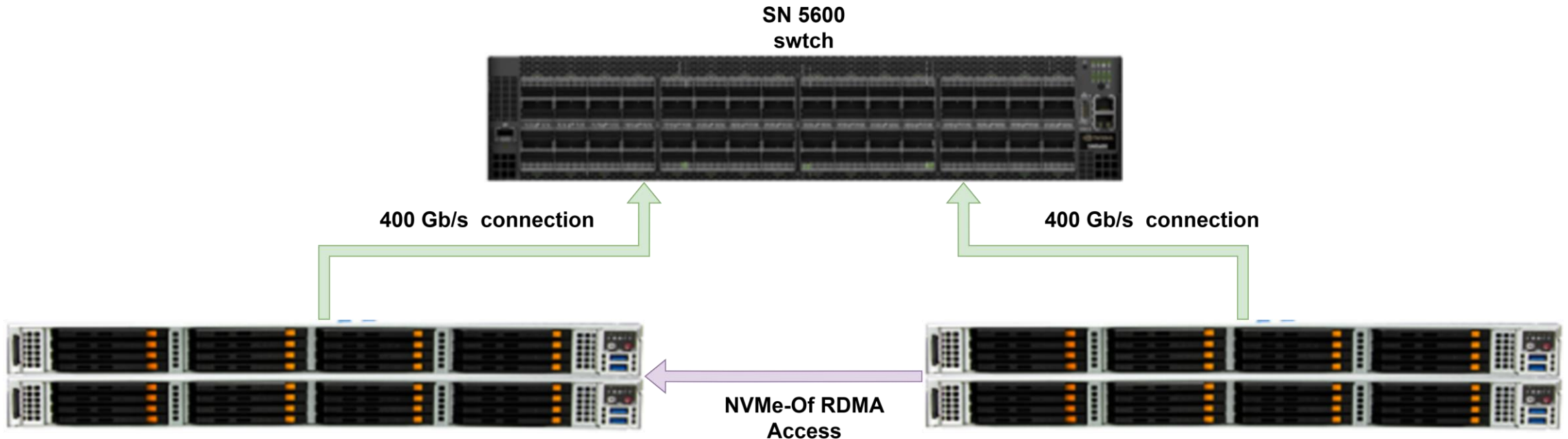
*By maximizing efficiency per core while maintaining native latency levels, xiRAID Opus helps organizations achieve their performance goals without the power and cost overhead of traditional high-core-count deployments.”*

**intel**

**Sonel Popatlal**

EMEA Leader – DC, AI, Client and Gigafactory, Intel

# xiRAID Opus performance measurements via 400Gb/s RoCE network and NVMe-oF protocol



## xiRAID Opus NVMe-Of target:

- **Platform:** Supermicro ASG-1115S-NE316R
- **xiRAID Opus 1.2:** RAID 6 with 7x Micron 9550 PRO (7.68TB)
- **OS:** Ubuntu 22.04

## xiRAID Opus NVMe-Of Initiator:

- **Platform:** Supermicro ASG-1115S-NE316R
- **OS:** Ubuntu 22.04

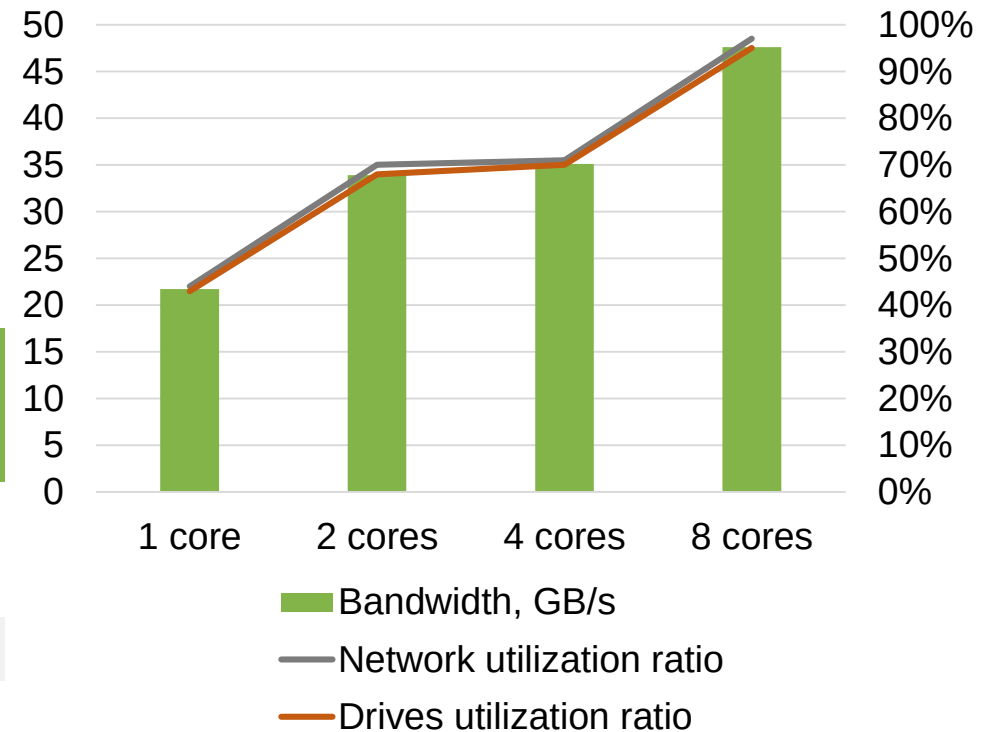
# Sequential tests on xiRAID Opus NVMe-of target from Linux NVMe-oF initiator

RAID 6 theoretical maximum sequential performance, based on Micron 9550 specification:

- Sequential Read: 14GB/s (one drive) \* 7 drives = **98GB/s**
- Sequential Write: is calculated taking in account penalties for RAID6 with 2 parity drives (5 drives \* 10GB/s) = **50 GB/s**
- Maximum Network throughput = 400Gb (**~49GB/s**)
- xiRAID Opus was tested with varying numbers of cores
- FIO workload pattern: 8 jobs, 32 iodepth

Pattern	Bandwidth	Network utilization ratio	Drives utilization ratio
<b>Sequential read: 1 core</b>	49.0GB/s	100%	(network limit)
<b>Sequential write: 1 core</b>	21.7GB/s	44%	43%
<b>Sequential write: 2 cores</b>	33.9GB/s	70%	68%
<b>Sequential write: 4 cores</b>	35.1GB/s	71%	70%
<b>Sequential write: 8 cores</b>	47.6GB/s	97%	95%

## Sequential write performance and latency



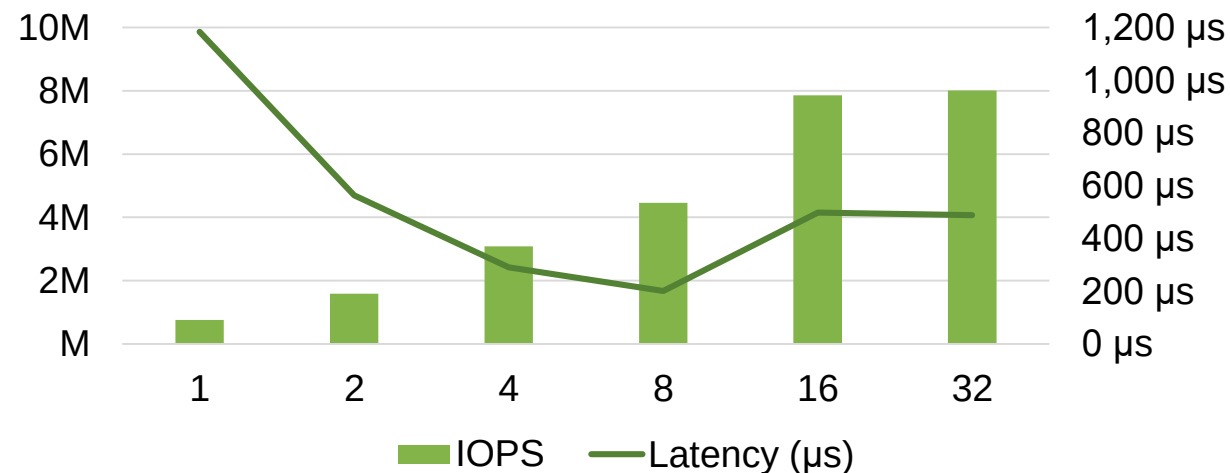
# Random read tests on xiRAID Opus NVMe-oF target from Linux NVMe-oF initiator

The RAID 6 theoretical random read performance maximum according to 400Gb/s network throughput:

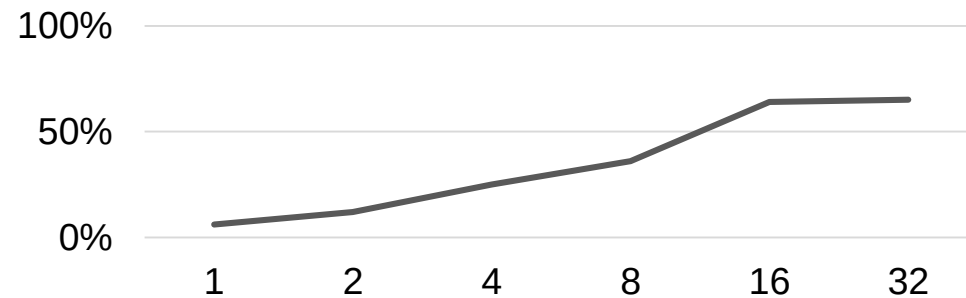
- Maximum Practical Network IOPS -> **8 M IOPS**
- xiRAID Opus was tested with varying numbers of cores

Pattern	IOPS	Latency (µs)	Network utilization theoretical ratio
<b>8 jobs 32 iodepth</b>			
<b>1 cores</b>	757k	1184.02	6%
<b>2 cores</b>	1590k	563.38	12%
<b>4 cores</b>	3086k	290.16	25%
<b>8 cores</b>	4454k	200.92	36%
<b>16 jobs 32 iodepth</b>			
<b>16 cores</b>	7859k	497.48	64%
<b>32 cores</b>	8010k	488.17	65%

## Random read performance and latency



## Random read performance drive usage ratio



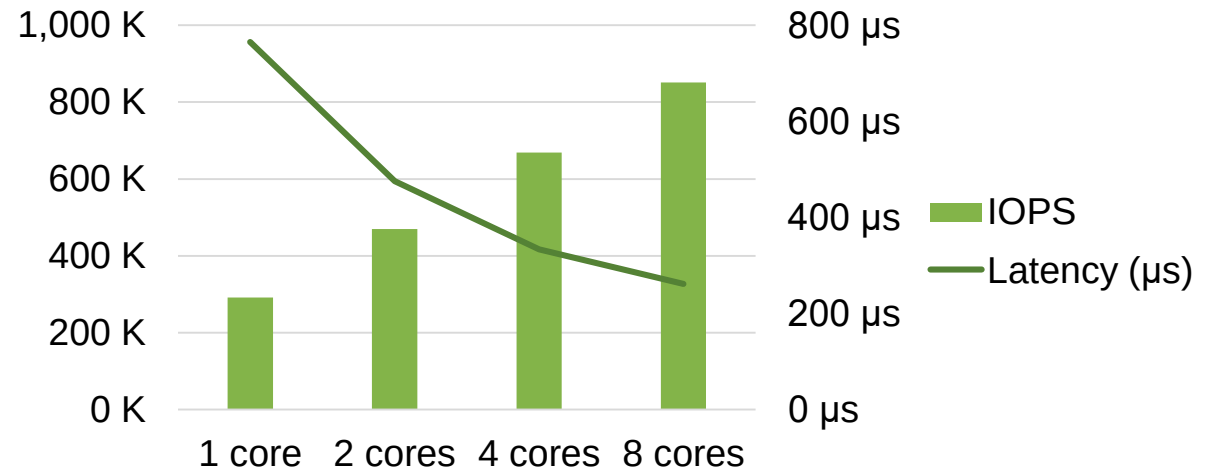
# Random write tests on xiRAID Opus NVMe-oF target from Linux NVMe-oF initiator

RAID 6 theoretical maximum random write performance, based on Micron 9550 specification:

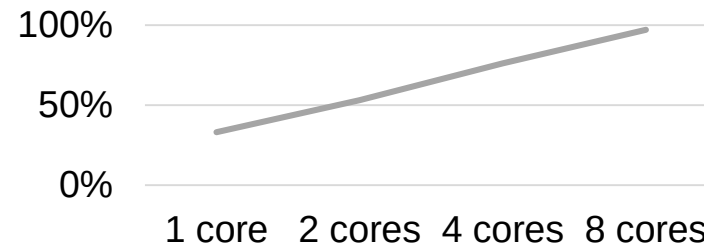
- Random write: 33%\* of raw performance due to Read-Modify-Writes penalty for RAID6: 380k IOPS (one drive) x 7 drives x 33% = **877k IOPS**
- xiRAID Opus was tested with multiple numbers of cores

Pattern	IOPS	Latency (µs)	Drives utilization ratio
1 core	292k	765.03	33%
2 cores	470k	475.52	53%
4 cores	669k	333.37	76%
8 cores	851k	261.86	97%

Random write performance and latency



Random write performance drive usage ratio



INNOR

# Roadmap



# The most efficient IO500 storage cluster

The Helma Cluster at NHR@FAU (leading German AI research center) is deployed with **xiRAID**, which makes the difference: *all other components are commodity hardware or open-source software:*

- 5PB Highly Available storage cluster serving 768 GPUs
- 3<sup>rd</sup> fastest IO500 storage cluster in production worldwide and the fastest for Lustre

IO500 is an industry most recognizable benchmark to compare performance of HPC and AI Storage Clusters

<https://io500.org/submissions/view/736>

## IO<sup>500</sup>

### Production ISC25 List

# ↑	INFORMATION				open-source				SCORE ↑
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	• The file system (Lustre 2.16.1) is also available as open-source				
1	SC23	Argonne National Laboratory	Aurora	Intel	• Xinnor xiRAID Classic (4.2.0) is proprietary software RAID solution requiring a purchased license				32,165.90
2	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo					2,508.85
3	ISC25	Erlangen National High Performance Computing Center	Helma	MEGWARE	Lustre	186	18,600	838.99	
4	ISC25	Samsung Electronics	SSC-24	WekaIO	WekaIO	291	16,005	826.86	
5	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	
6	SC24	MSKCC	IRIS	WekaIO	WekaIO	261	27,144	665.49	
7	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	
8	SC24	SoftBank Corp	CHIE-3	DDN	EXAScaler	240	26,880	500.20	
9	ISC25	Joint Center for Advanced High Performance Computing	Miyabi-G	DDN	Lustre	200	1,600	391.60	
10	SC24	Danish Centre for AI innovation AS	GEFION	DDN	EXAScaler	128	12,288	368.56	

The software stack includes both open-source and proprietary components:

- The operating system (AlmaLinux 9.4) is available as open-source
- The file system (Lustre 2.16.1) is also available as open-source
- Xinnor xiRAID Classic (4.2.0) is proprietary software RAID solution requiring a purchased license

# Xinnor exploits the full hardware potential

IO<sup>500</sup>



Comparing 2 clusters on the IO500 top 10, both based on:

- Celestica SC6100 server (branded DDN ES400NVX in the DDN deployment).
- The same parallel file system: Lustre

#	System	Solution (Vendor)	IO500 Score	Number of systems	IO500 Score / storage systems	Number of NVMe drives	IO500 Score / drive
3	Helma	xiRAID + Lustre (Xinnor)	838.99	10	83.9	240	3.5
7	Leonardo*	ExaScaler (DDN)	648.96	29	22.4	688	0.9
*On top of 29 NVMe systems, the Leonardo cluster also includes 180 HDD for OST				1/3	4x	1/3	3.7x

# Helma Storage Cluster at NHR@FAU

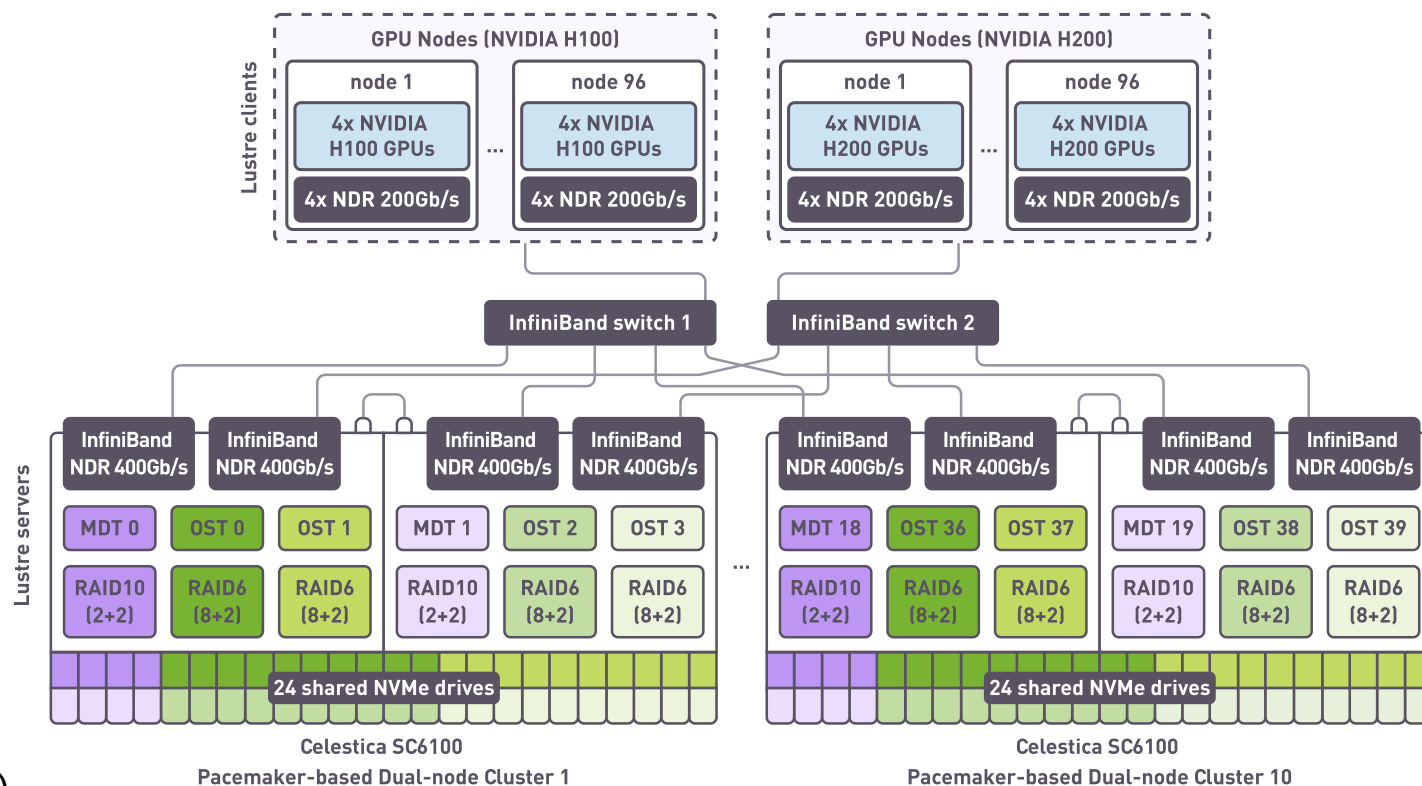
IO<sup>500</sup>

5PB HA storage cluster to serve 768 GPUs

IOR & FIND		METADATA	
EASY WRITE	811.33 GiB/s	EASY WRITE	1,819.16 kiOP/s
EASY READ	1,798.77 GiB/s	EASY STAT	8,221.83 kiOP/s
HARD WRITE	60.52 GiB/s	EASY DELETE	1,420.24 kiOP/s
HARD READ	419.04 GiB/s	HARD WRITE	387.63 kiOP/s
FIND	3,017.00 kiOP/s	HARD READ	2,236.33 kiOP/s
		HARD STAT	3,358.07 kiOP/s
		HARD DELETE	235.84 kiOP/s

10 Celestica SC6100 SBB servers, each made of:

- AMD EPYC 9454P
- 384GB DDR5 RAM
- 4x Phison Pascari PCIe5 X200E 6.4TB (3DWPD)
  - 2x 2+2 namespaces xiRAID RAID10
- 20x Phison Pascari PCIe5 X200P 30.72TB (1DWPD)
  - 4x 8+2 namespaces xiRAID RAID6 128K stripe size
- 2x NVIDIA CX-7 NDR400



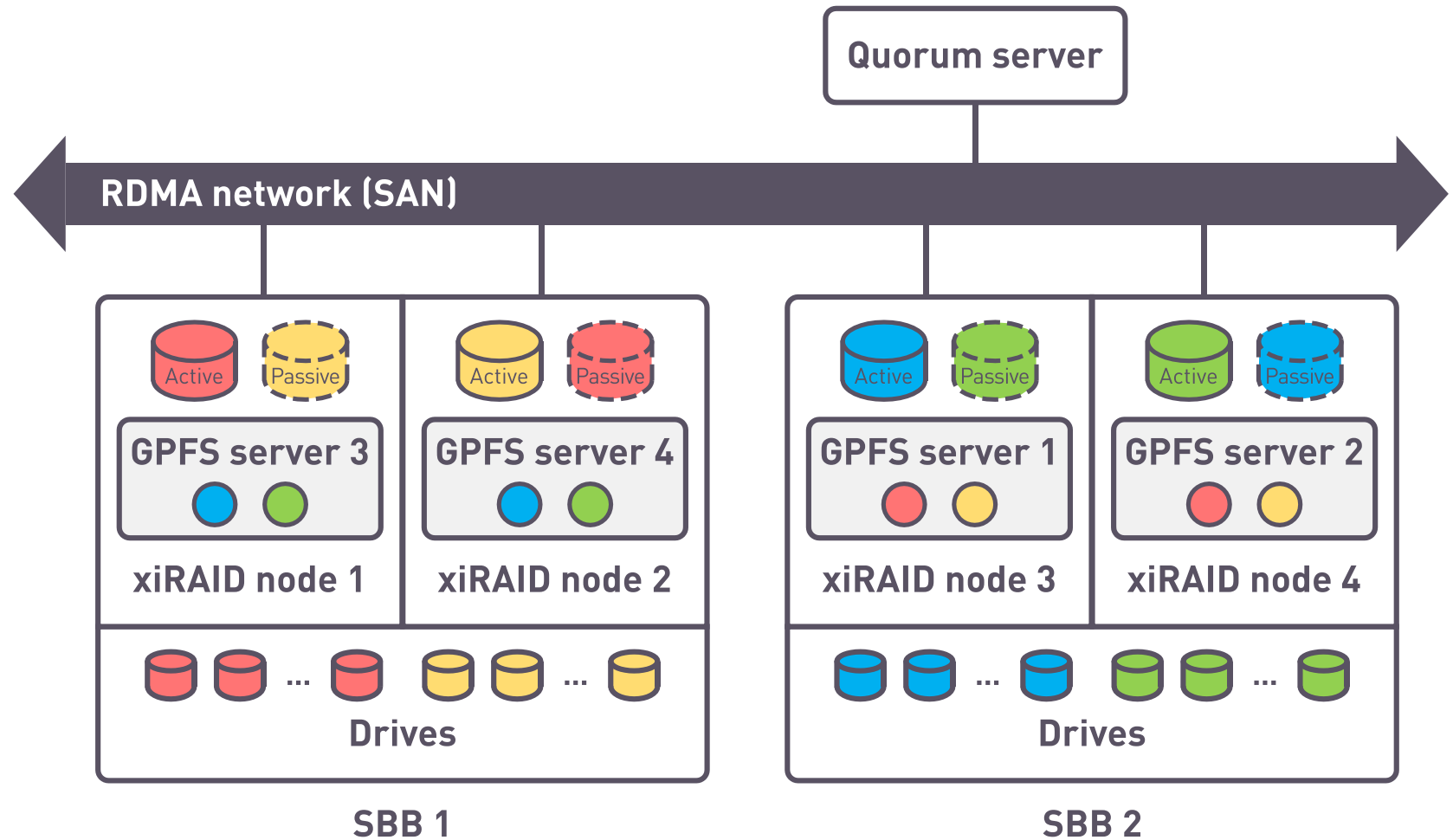
# SBB systems

- Two servers and set of dual-ported NVMe drives in a single box
- Each server has access to all the drives
- Some models have internal network connection between servers
- All servers have IPMI for fencing
- Available from multiple vendors



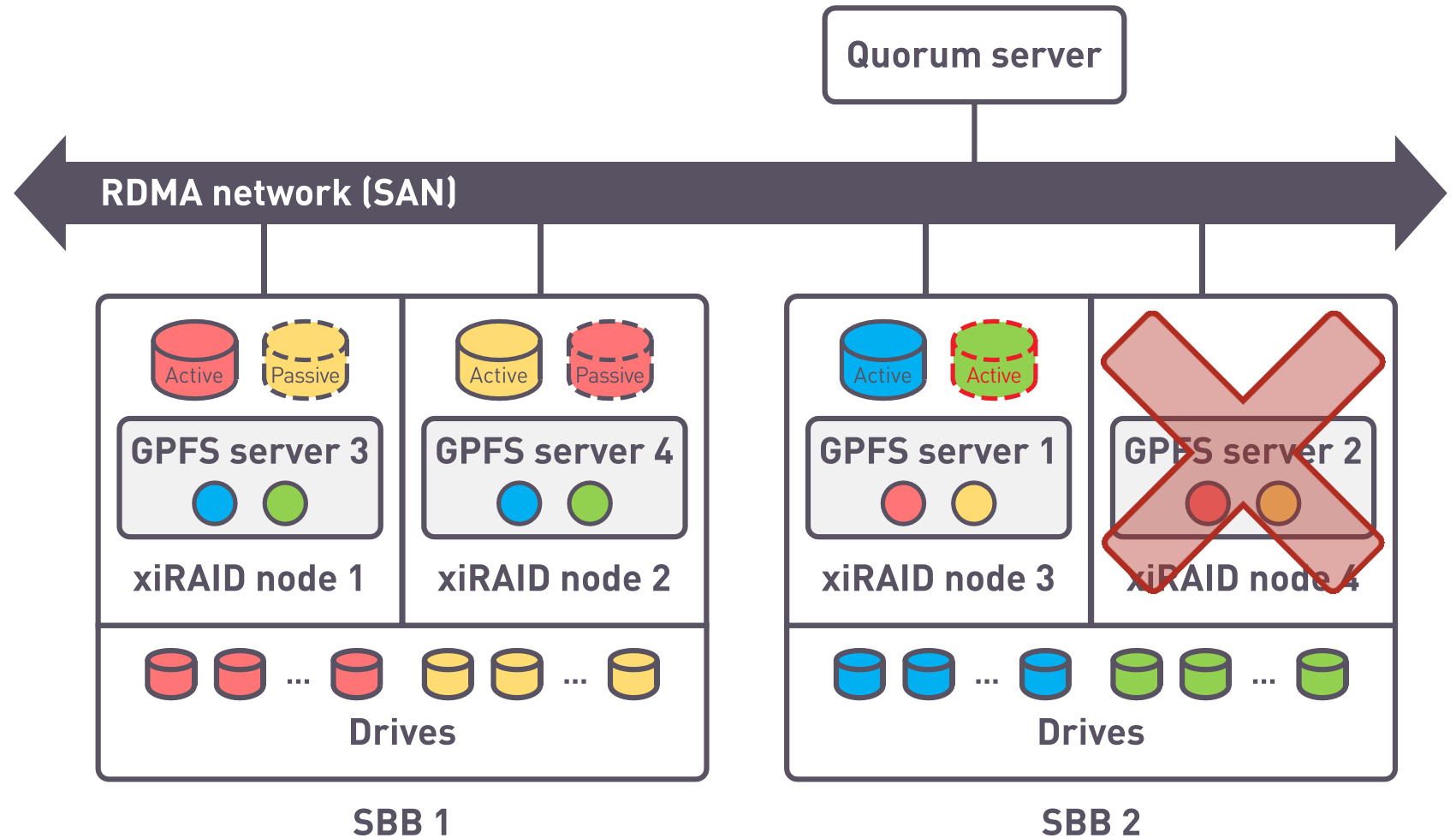
## 2 convergent SBBs

**Main drawback:** all user data is transmitted over the network twice: from a client to a NSD server and from the NSD server to a block storage (NVMeOF).



## 2 convergent SBBs: failover

- RAID is failovered to Node #3, changing ANA status on Node #3
- Multipath on xiRAID node #1 and #2 changes it's configuration, allowing GPFS server #3 and #4 continue to work with RAID.
- GPFS server #1 will failover the GPFS server #2 workload (GPFS mechanism)



# Roadmap

Dual controller cluster support with:

- RAID clustering;
- NVMeOF RDMA target clustering;
- ANA support for multipath on the initiator side.

GA: H1 2027

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

