VALIDATED DESIGN

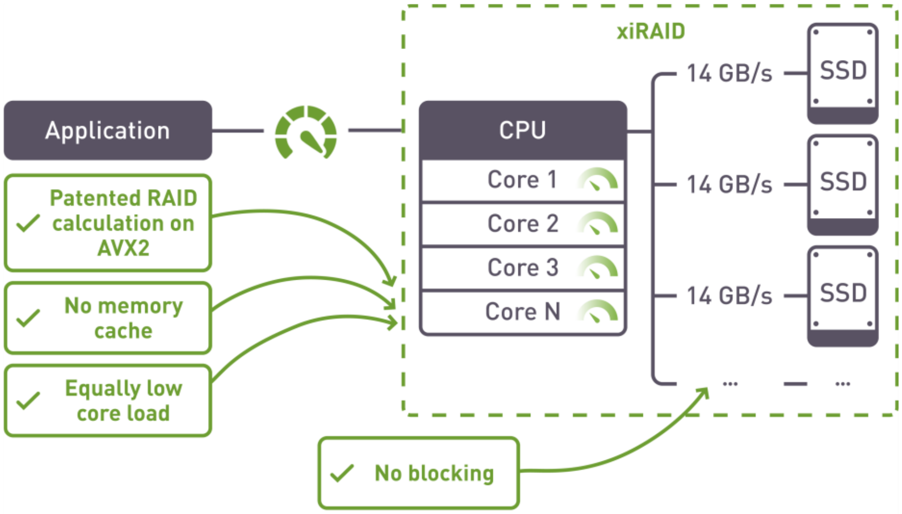
IBM STORAGE SCALE WITH SUPERMICRO SERVERS AND XINNOR XIRAID

TABLE OF CONTENTS

Executive Summary

As AI/ML, Agentic AI, and Retrieval-Augmented Generation (RAG) workloads grow in complexity and demand, the performance of underlying storage systems becomes mission-critical. To keep GPUs fully utilized, storage must deliver exceptional performance in both sequential and random operations. With their excellent performance, NVMe PCIe drives are ideal for mission-critical applications.

This reference architecture demonstrates a high-performance, highly available, and cost-effective storage cluster built with IBM Storage Scale (formerly GPFS), Xinnor's xiRAID software RAID, and Supermicro NVMe storage servers. The solution combines xiRAID's RAID6 protection and parallel performance optimization with IBM Storage Scale's distributed data services to deliver scalable storage that meets the performance and reliability demands of AI and HPC environments.



By combining efficient checksum calculation using x86 CPU's AVX (Advanced Vector Extensions) instruction set technology with Xinnor's proprietary lockless data path, xiRAID achieves near-raw drive performance with minimal system resource utilization. AVX is an x86 CPU vector instruction where a single command executes operations on multiple values simultaneously, processing larger amounts of data than scalar CPU instructions. Lockless data path is Xinnor's method to guarantee no race conditions between multiple write requests to a single stripe, by ensuring all write operations for a stripe execute on a single worker thread, avoiding traditional stripe locking. The combination of these two technologies improves checksum calculation speed, minimizes CPU utilization, reduces latency, and increases throughput.

xiRAID supports multiple RAID levels, including RAID 0, 1, 5, 6, RAID 7.3 (3 drives parity), all nested RAIDs (10, 50, 60, 70), and N+M. This flexibility enables customers to select their preferred protection level, tailored to the specific application requirements. LED and email support allow easy system integration and manageability.

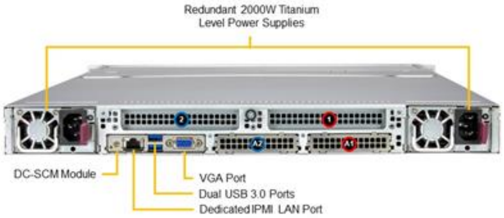
xiRAID is in production at leading research institutions, and its superior performance is demonstrated by its appearance in the prestigious IO500 list.

IBM Storage Scale

IBM Storage Scale is a software-defined file and object storage solution that enables a global data platform for artificial intelligence (AI), high-performance computing (HPC), advanced analytics, and other demanding workloads. It seamlessly manages unstructured data, including documents, audio, images, and videos, across distributed storage environments.

Built on a massively parallel file system, it provides global data abstraction and unified access across x86, IBM Power, IBM Z®, ARM-based POSIX clients, virtual machines, and Kubernetes, even in non-IBM storage environments.

IBM Storage Scale (formerly called IBM Spectrum Scale and GPFS) offers multiple licensing models (editions), each tailored for different use cases and levels of functionality. Each IBM Scale edition supports various data protection schemes, ranging from local replication (without any protection at the drive level) to geo-dispersed erasure coding. The most popular edition is the Data Management Edition (DME). In the DME, server node protection is ensured via a Replication Factor (RF=2, RF=3). Strong consistency must be used to provide data protection when JBOD/JBOF media is used. Write performance can be improved by combining replication with eventual consistency and RAID-protected media. This task can be accomplished by Xinnor's xiRAID, guaranteeing data durability and performance.



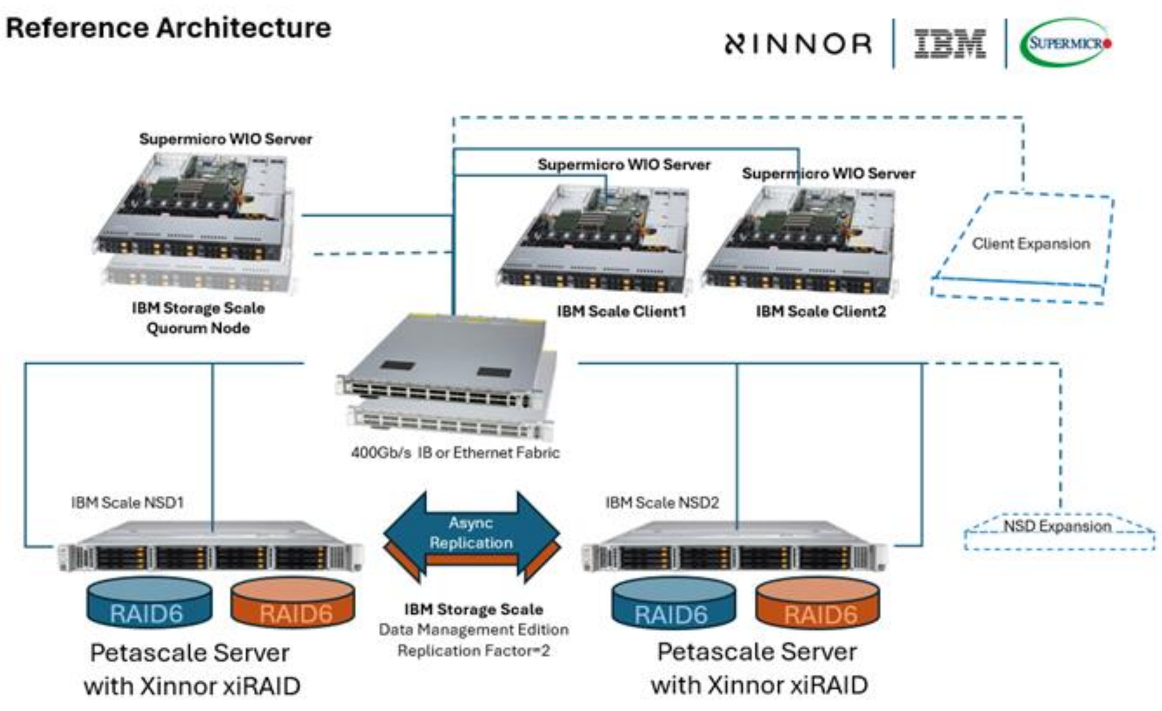
	Slot Description
1 2	PCIe 5.0 x16 FHHL
A1 A2	PCIe 5.0 x16 AIOM

Supermicro’s Petascale Storage Servers

Supermicro's Petascale All-Flash servers are designed for large-scale and performance-intensive applications. The system balances PCIe lanes across storage media and network I/O, ensuring non-blocking bandwidth for network-based clients. The optimized thermal design, in accordance with EDSFF specifications, contributes to efficient operation and reduced energy consumption.

Customers have the choice of three platforms (Dual Intel Xeon 6700 series, NVIDIA Grace Superchip, or Single AMD EPYC 9004/9005 series), available in 1U and 2U configurations, reaching storage capacities up to 1.9 PB using E3.S media. The system also boasts exceptional internal PCIe 5.0 bandwidth, perfect for demanding workloads.

Component	Technology	Role & Benefit
Server Hardware	Supermicro NVMe Servers	High-density, cost-efficient, software-defined storage platform
Drive Protection	xiRAID RAID6	Fast, CPU-efficient software RAID with up to 2-drive failure tolerance
Filesystem & Cluster	IBM Storage Scale (GPFS)	A parallel distributed filesystem with replication and unified access, which creates Network Shared Disks (NSD) to clients writing to a filesystem with directory-based file sets that can include their own inode space for special data placement and management.



1. Hardware Layer

- *Data Nodes:* Supermicro SSG-122B-NE316R-1 / -2
  - CPU: Intel Xeon 6756E (128 cores) –*for xiRAID+IBM Scale 32 cores are sufficient*
  - Memory: 16× 64GB DDR5-6400 ECC RDIMM – *to note that for xiRAID+IBM Scale, a total of 256GB DRAM is sufficient as long as all memory DIMM slots are populated.*
  - Storage: 6× SK Hynix PS1010 15.36TB NVMe PCIe Gen5x4 SSDs
  - Networking: 2x Mellanox ConnectX-7
- *Control and Client Nodes:*
  - Supermicro AS-1114S-WN10RT-1 (Quorum)
  - Supermicro AS-1114S-WN10RT-2/-3 (Client Nodes)
    - Networking: 1x Mellanox ConnectX-7

2. System Software:

OS: RHEL 9.4
Kernel Version: 5.14.0-427.42.1.19_4. x86_64
MOFED Version: MLNX_OFED_LINUX-24.10-2.1.8.0-rhe19.4-x86_64

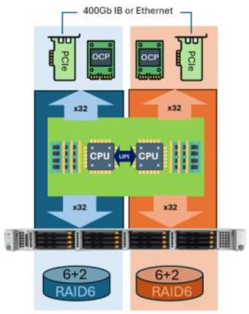
3. Drive-Level Protection

- **Technology**: Xinnor xiRAID in RAID6
- **Function**: Software-based block-level RAID
- **Benefit**: Protects against failures of up to 2 drives in each server

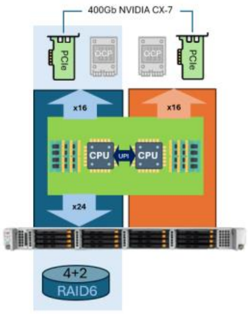
4. Multi-node-Level Protection

- **Technology**: IBM Storage Scale (GPFS) based on 5.2.0 or later
- **Function**: Distributed, parallel POSIX-compliant filesystem
- **Data Protection**: NSD (Network Shared Disks) Replication Factor (RF) (e.g., RF=2 or RF=3) for node/server-level redundancy
  - Note that best practice calls for metadata to be a RF= 2, and specific data can be mapped to a directory fileset can default to RF of 1, 2, or 3 in the filesystem. In this case, we implemented RF-2 on both data and metadata

5. Reference Architecture and POC NSD Configuration



Optimal NSD Configuration



Tested NSD Configuration

Validation Workflow

1 – Testbed Preparation

Cluster layout: two Supermicro SSG-122B-NE316R NSD servers, each with 6 × 15.36 TB PCIe 5.0 NVMe SSDs combined into a xiRAID RAID-6 (4+2) set; client and quorum nodes connected via 200 Gb InfiniBand (ConnectX-7).

1.1) MOFED Installation

1.2) xiRAID Classic 4.2 Installation

1.3) IBM Software Installation across all nodes

2 – Block-Layer Performance (xiRAID)

Goal: to prove that each xiRAID array delivers ≥ 85 % of the raw-SSD bandwidth.

Parameter	fio Options
Sequential	--rw=write/read --bs=1M --iodepth=32 --numjobs=4
Random 4 K	--rw=randread --bs=4k --iodepth=8 and iodepth=1

Acceptance thresholds

Based on the drive specification and RAID6 over-head, the acceptable thresholds are:

- seq_write ≥ 40 GB/s
- seq_read ≥ 80 GB/s

3 – File-System Performance (IBM Storage Scale)

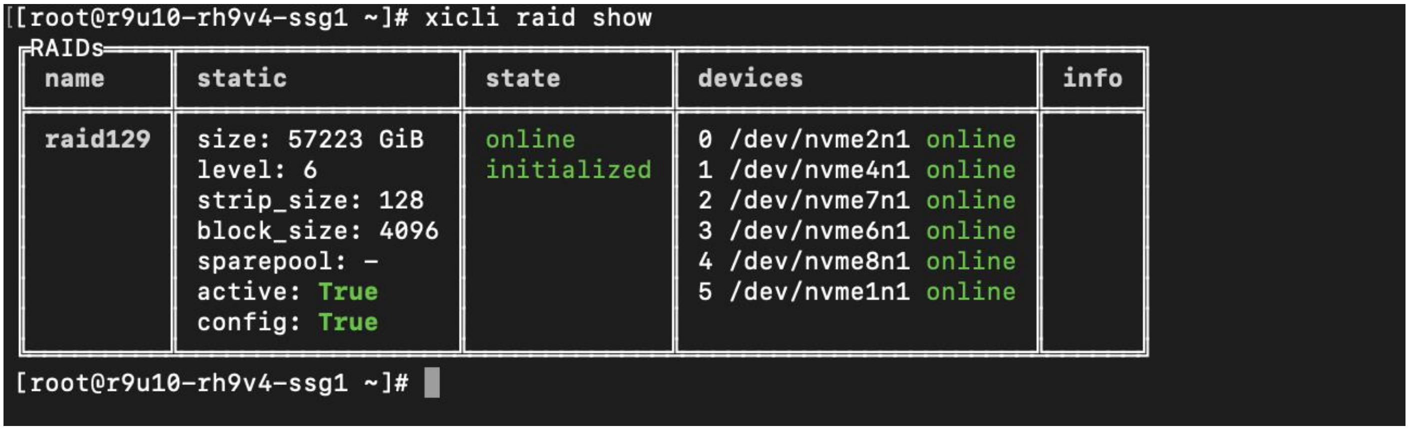
Goal: to confirm that xiRAID introduces ≤ 15% overhead relative to IBM Storage Scale and ensure maximum bandwidth is achievable from the client network adapters.

4 – Fault-Tolerance Test

Goal: to verify uninterrupted service when an entire IBM Storage Scale Network Shared Disk (NSD) server fails.

Deployment Details

xiRAID was installed on the Supermicro servers to create fault-tolerant, high-performance block devices. These RAID groups were presented as local storage volumes to IBM Storage Scale. The Storage Scale NSDs were created on top of the xiRAID-protected volumes. A GPFS filesystem (gpfs0) was mounted using these NSDs.



```
[[root@r9u10-rh9v4-ssg1 ~]# xicli raid show
 ┌RAIDs─────────────────────────────────────────────────────────────────────────┐
 │ name    │ static              │ state        │ devices                │ info │
 │─────────┼─────────────────────┼──────────────┼────────────────────────┼──────│
 │ raid129 │ size: 57223 GiB     │ online       │ 0 /dev/nvme2n1 online  │      │
 │         │ level: 6            │ initialized  │ 1 /dev/nvme4n1 online  │      │
 │         │ strip_size: 128     │              │ 2 /dev/nvme7n1 online  │      │
 │         │ block_size: 4096    │              │ 3 /dev/nvme6n1 online  │      │
 │         │ sparepool: -        │              │ 4 /dev/nvme8n1 online  │      │
 │         │ active: True        │              │ 5 /dev/nvme1n1 online  │      │
 │         │ config: True        │              │                        │      │
 └──────────────────────────────────────────────────────────────────────────────┘
[root@r9u10-rh9v4-ssg1 ~]#
```

RAID groups were configured with default parameters.

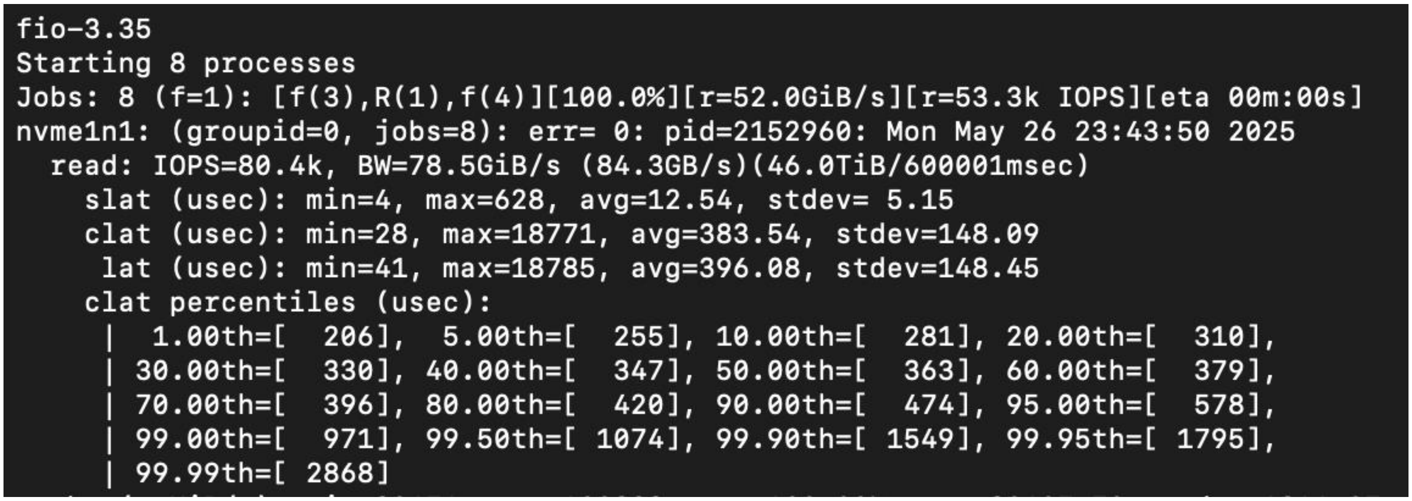
The RAID performance

	Raw Drive	RAID6 theoretical max performance	xiRAID RAID6 measured performance	Efficiency
Sequential Read (GB/s)	14,5	87,0	84,3	97%
Sequential Write (GB/s)	9,3	37,2	36,8	98,4%
Random Read (KIOPS)	1021*	6126	6125	100%
Random Write (KIOPS)	485	970**	971	100%

*Measured performance

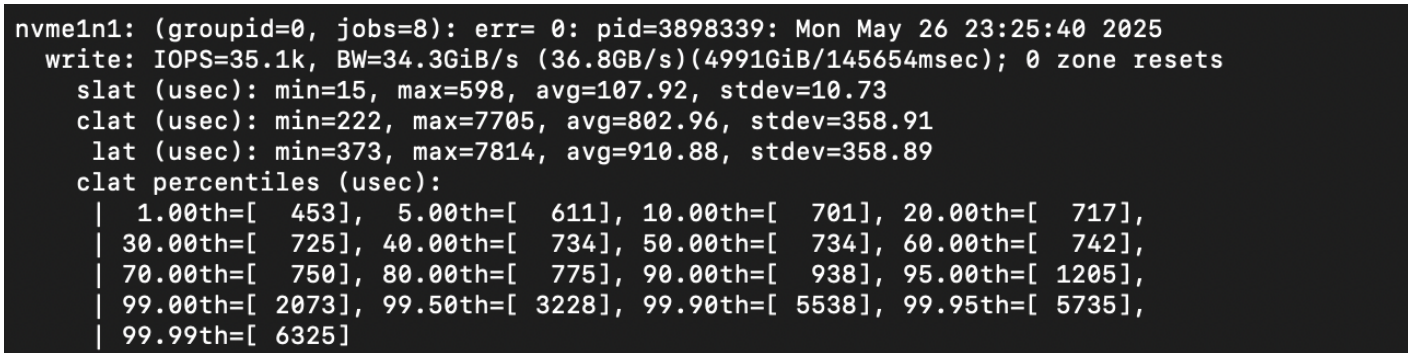
**In RAID 6, the max theoretical performance is about 33% of the raw drive performance due to read-modify-write operations and the overhead of the parity calculations: each small random write typically involves 3 write operations (1 for data and 2 for parity) and 3 read operations (to fetch the old data and parity).

Sequential read screenshot (8 threads 4 QD 1M block size):



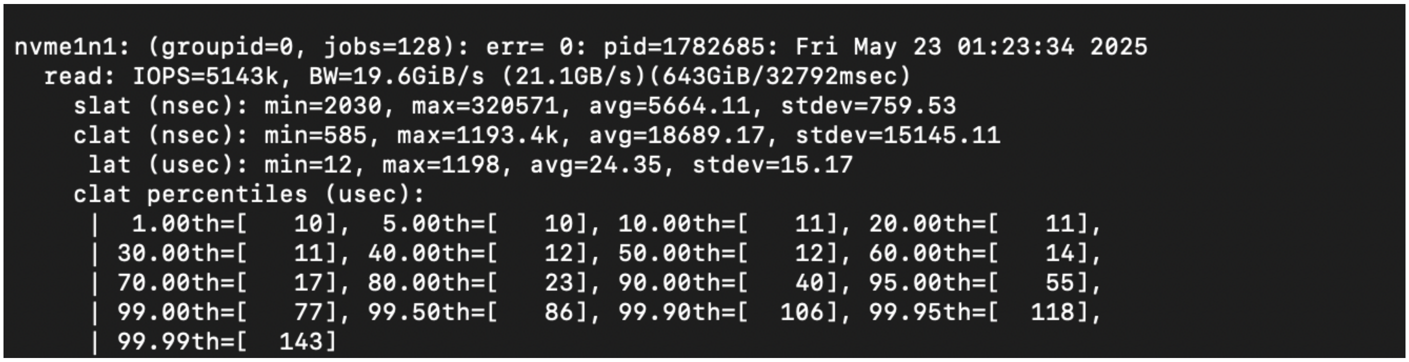
```
fio-3.35
Starting 8 processes
Jobs: 8 (f=1): [f(3),R(1),f(4)][100.0%][r=52.0GiB/s][r=53.3k IOPS][eta 00m:00s]
nvme1n1: (groupid=0, jobs=8): err= 0: pid=2152960: Mon May 26 23:43:50 2025
  read: IOPS=80.4k, BW=78.5GiB/s (84.3GB/s)(46.0TiB/600001msec)
    slat (usec): min=4, max=628, avg=12.54, stdev= 5.15
    clat (usec): min=28, max=18771, avg=383.54, stdev=148.09
     lat (usec): min=41, max=18785, avg=396.08, stdev=148.45
    clat percentiles (usec):
     |  1.00th=[  206],  5.00th=[  255], 10.00th=[  281], 20.00th=[  310],
     | 30.00th=[  330], 40.00th=[  347], 50.00th=[  363], 60.00th=[  379],
     | 70.00th=[  396], 80.00th=[  420], 90.00th=[  474], 95.00th=[  578],
     | 99.00th=[  971], 99.50th=[ 1074], 99.90th=[ 1549], 99.95th=[ 1795],
     | 99.99th=[ 2868]
```

Sequential write screenshot (8 threads 4 QD 1M block size):



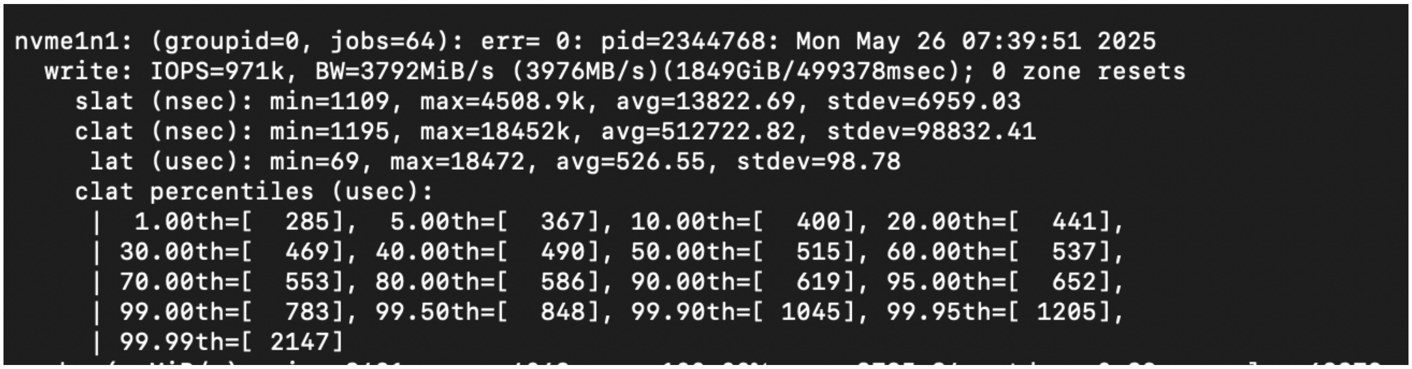
```
nvme1n1: (groupid=0, jobs=8): err= 0: pid=3898339: Mon May 26 23:25:40 2025
  write: IOPS=35.1k, BW=34.3GiB/s (36.8GB/s)(4991GiB/145654msec); 0 zone resets
    slat (usec): min=15, max=598, avg=107.92, stdev=10.73
    clat (usec): min=222, max=7705, avg=802.96, stdev=358.91
     lat (usec): min=373, max=7814, avg=910.88, stdev=358.89
    clat percentiles (usec):
     |  1.00th=[  453],  5.00th=[  611], 10.00th=[  701], 20.00th=[  717],
     | 30.00th=[  725], 40.00th=[  734], 50.00th=[  734], 60.00th=[  742],
     | 70.00th=[  750], 80.00th=[  775], 90.00th=[  938], 95.00th=[ 1205],
     | 99.00th=[ 2073], 99.50th=[ 3228], 99.90th=[ 5538], 99.95th=[ 5735],
     | 99.99th=[ 6325]
```

Random Read screenshot



```
nvme1n1: (groupid=0, jobs=128): err= 0: pid=1782685: Fri May 23 01:23:34 2025
  read: IOPS=5143k, BW=19.6GiB/s (21.1GB/s)(643GiB/32792msec)
    slat (nsec): min=2030, max=320571, avg=5664.11, stdev=759.53
    clat (nsec): min=585, max=1193.4k, avg=18689.17, stdev=15145.11
     lat (usec): min=12, max=1198, avg=24.35, stdev=15.17
    clat percentiles (usec):
     |  1.00th=[   10],  5.00th=[   10], 10.00th=[   11], 20.00th=[   11],
     | 30.00th=[   11], 40.00th=[   12], 50.00th=[   12], 60.00th=[   14],
     | 70.00th=[   17], 80.00th=[   23], 90.00th=[   40], 95.00th=[   55],
     | 99.00th=[   77], 99.50th=[   86], 99.90th=[  106], 99.95th=[  118],
     | 99.99th=[  143]
```

Random write Screenshot - Libaio (QD=8):



```
nvme1n1: (groupid=0, jobs=64): err= 0: pid=2344768: Mon May 26 07:39:51 2025
  write: IOPS=971k, BW=3792MiB/s (3976MB/s)(1849GiB/499378msec); 0 zone resets
    slat (nsec): min=1109, max=4508.9k, avg=13822.69, stdev=6959.03
    clat (nsec): min=1195, max=18452k, avg=512722.82, stdev=98832.41
     lat (usec): min=69, max=18472, avg=526.55, stdev=98.78
    clat percentiles (usec):
     |  1.00th=[  285],  5.00th=[  367], 10.00th=[  400], 20.00th=[  441],
     | 30.00th=[  469], 40.00th=[  490], 50.00th=[  515], 60.00th=[  537],
     | 70.00th=[  553], 80.00th=[  586], 90.00th=[  619], 95.00th=[  652],
     | 99.00th=[  783], 99.50th=[  848], 99.90th=[ 1045], 99.95th=[ 1205],
     | 99.99th=[ 2147]
```

Replication Setup

A directory file set within the filesystem **gpfs0** was configured with a Replication Factor (RF) of 2, enabling GPFS to store two copies of each data block across different NSDs.

In the tested configuration, the maximum possible filesystem performance is limited by the number of NVMe devices in the NSD, the number of clients, and the performance of the network adapters on the client servers, which is 2 x 200 Gbps. To measure even higher performance, we recommend equipping the client servers with a higher-speed network or adding more clients to the cluster.

Reliability Validation

To validate system resiliency, one of the NSD servers (NSD2) was rebooted during testing. While NSD2 was offline, the system remained fully operational: data reads and writes continued uninterrupted. Upon reboot, the NSD server rejoined the cluster, and GPFS automatically re-integrated the affected disks. No manual intervention was required; all processes functioned as expected due to the combination of xiRAID's local disk protection and IBM Storage Scale's replication at the filesystem level.

Reference Architecture at Scale

This two-node configuration demonstrates that a minimal configuration can deliver exceptional performance and value. The cluster can scale both vertically by adding more drives per server and horizontally by adding more NSD nodes and clients to the cluster.

Case Study

A production deployment at Aalen University implemented this architecture designed by ABC systems AG, using:

- 2 Supermicro server nodes
- 10× 7.68TB Kioxia CD8P Gen5 SSDs per node
- xiRAID (RAID6) + IBM Storage Scale

Details of that architecture can be found in this case study https://xinnor.io/case-studies/aalen/

The case study demonstrates how it is possible to build a highly resilient Storage cluster using xiRAID, IBM Scale, and Supermicro Servers.

In that specific deployment, 10 x 7.68TB NVMe PCIe 5.0 Kioxia CD8P KCD8XPUG7T68 were used in each of the 2 servers and protected by xiRAID in RAID6.

The results of the tests performed on each of the servers showed extraordinary performance, close to the theoretical maximum of the underlying hardware:

```
GPFS FS with gpfsperf:

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

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

Performance summary from each server node at Aalen University:

	Raw Drive Capability without RAID and a file system	Measured (GPFS)	Efficiency
Sequential Write (GB/s)	55,0	42.8	78%*
Sequential Read (GB/s)	92,0	80.2	87%

*The efficiency is measured over 10 drives, including the parity drives. If we exclude the parity drives, the write efficiency is 97%

Conclusion

This reference design, along with the use case at Aalen University, demonstrates that enterprise-grade storage for AI and HPC workloads can be built using industry-standard servers paired with innovative software-defined RAID and parallel file system technologies. The combination of xiRAID and IBM Storage Scale on Supermicro hardware delivers:

- Scalable, fault-tolerant software-defined storage
- Top-tier throughput and latency performance
- A reliable foundation for next-generation AI/ML applications.

The successful test of node reboot with replication factor 2 confirms the design's resilience and suitability for enterprise-scale workloads demanding high availability and performance.

Feature	Technology	Value Delivered
Dual-Layer Redundancy	xiRAID + GPFS	Protection from drive and node failures simultaneously
High Performance EDSFF NVMe Technology	Petascale Storage Servers with PCIe 5.0 Architecture	Software-defined server hardware eliminates barriers to entry
High Efficiency	xiRAID RAID6	90%+ hardware utilization under real-world workloads
Parallel Access	IBM Storage Scale	Optimized for GPU-intensive and I/O-demanding environments

Further Information

https://www.supermicro.com/en/products/nvme

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Learn more at www.supermicro.com

XINNOR

Traditional RAID implementations were designed for slow SATA and SAS disk drive storage media. With the adoption of NVMe drives, traditional hardware RAID becomes the bottleneck, as it operates via a 16-lane PCIe bus. Since each NVMe drive uses 4 PCIe lanes, hardware RAID can only address four drives at full speed. To overcome this limitation, Xinnor developed xiRAID, an innovative software RAID engine specifically designed for NVMe drives.

Learn more at: www.xinnor.com