

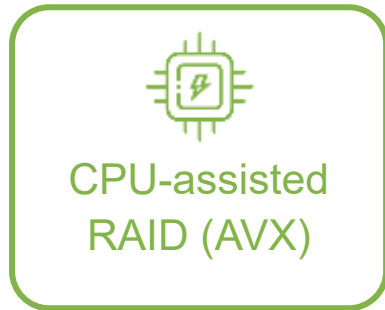
XINNOR

xiRAID + IBM Scale + Supermicro HW

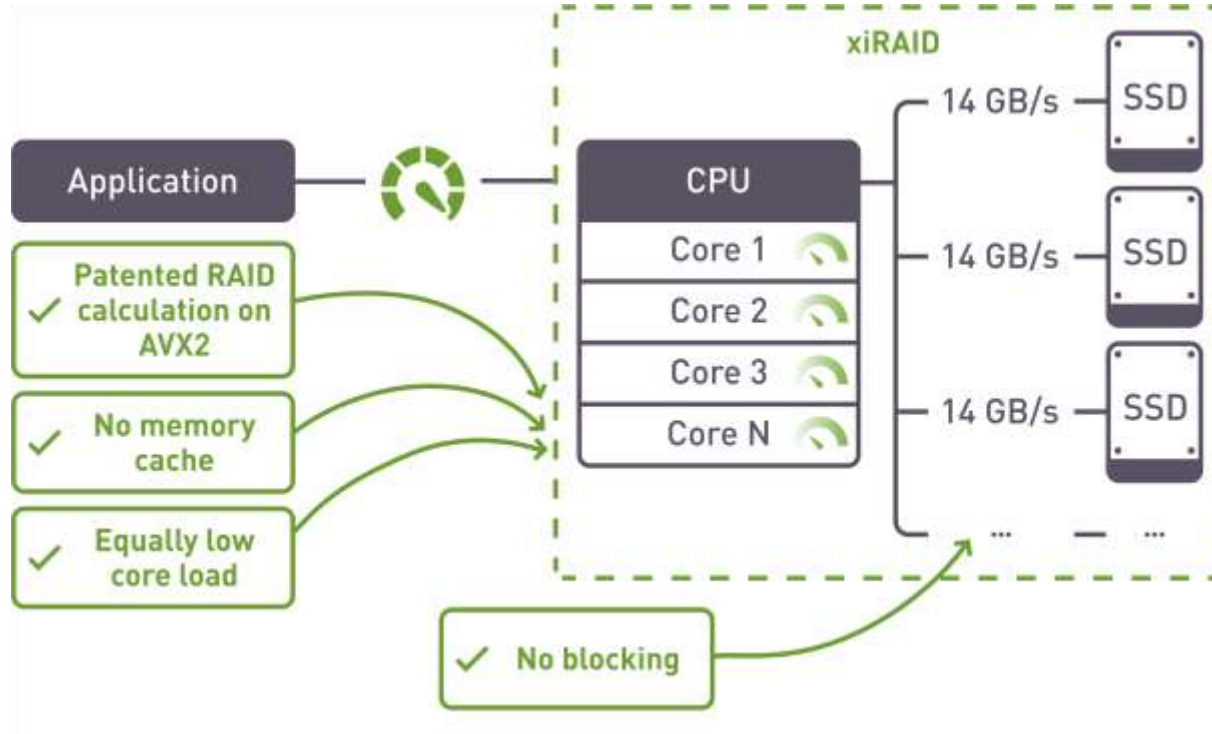
IBM Scale User Group, Aug '25
Davide Villa
davide.villa@xinnor.io



xiRAID Classic



Improved checksum
calculation speed.



Lower CPU utilization.



Reducing latency



Increasing throughput

Helma Storage Cluster at NHR@FAU

5PB HA storage cluster to serve 768 GPUs

Production ISC25 List

[Customize](#)[Download](#)**Production**

10 Node Production



Research



10 Node Research

Full

Historical

Ranking of production system submissions. This is a subset of the Full List of submissions, showing only one highest-scoring result per storage system. Submitters who want a submission that is currently on the Research List to be on the Production List should contact the IO500 Steering Committee.

#	INFORMATION							IO500			
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GiB/s)	MD (KIOP/s)	REPRO.
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	62,400	32,165.90	10,066.09	102,785.41	✓
2	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo	DAOS	90	6,480	2,508.85	742.90	8,472.60	✓
3	ISC25	Erlangen National High Performance Computing Center	Helma	MEGWARE	Lustre	180	18,600	838.99	438.62	1,604.84	✓

IOR & FIND

EASY WRITE	811.33 GiB/s
EASY READ	1,798.77 GiB/s
HARD WRITE	60.52 GiB/s
HARD READ	419.04 GiB/s
FIND	3,017.00 KIOP/s

METADATA

EASY WRITE	1,819.16 KIOP/s
EASY STAT	8,221.83 KIOP/s
EASY DELETE	1,420.24 KIOP/s
HARD WRITE	387.63 KIOP/s
HARD READ	2,236.33 KIOP/s
HARD STAT	3,358.07 KIOP/s
HARD DELETE	235.84 KIOP/s



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>

The background of the slide is a dark, abstract composition. It features a faint, semi-transparent image of a violin, positioned diagonally from the bottom left towards the top right. Overlaid on this are vibrant, ethereal light streaks and waves in shades of green, yellow, and orange, creating a sense of motion and energy. The overall aesthetic is high-tech and artistic.

xINNOV

xiRAID + IBM Scale + Supermicro Reference Architecture

Node-level data protection options for IBM Scale

- ☐ **IBM GPFS Native RAID (GNR).** Only available on proprietary hardware.
- ☐ **IBM Erasure Code Edition (ECE).** Minimum 4 NVMe storage nodes, with 6 being preferable.

Third-party RAID engines:

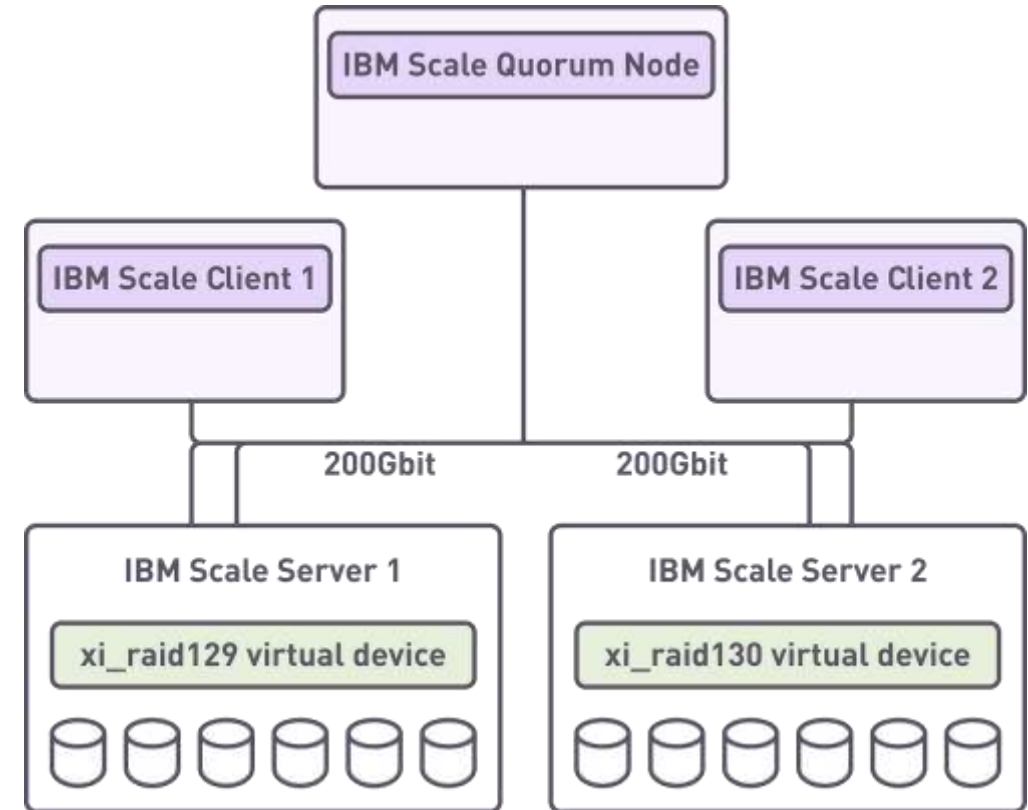
- ☐ **Hardware RAID** would have limited performance and wasted 16 precious PCIe lanes, forcing to add a switch (latency)
- ☒ **xiRAID Software RAID** protect data at max speed

Objective of the RA

- Enable IBM Scale to run on off-the-shelves NVMe servers to:
 - Fully utilize the performance of the hardware
 - Assure data integrity
 - Minimize cost
- Ideal solution for mid-size AI storage needs, where performance and cost-optimization matter

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 9.4
- 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



Deployment Details

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

```
[[root@r9u10-rh9v4-ssg1 ~]# xicli raid show
```

RAIDs				
name	static	state	devices	info
raid129	size: 57223 GiB	online initialized	0 /dev/nvme2n1	online
	level: 6		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 ~]#
```

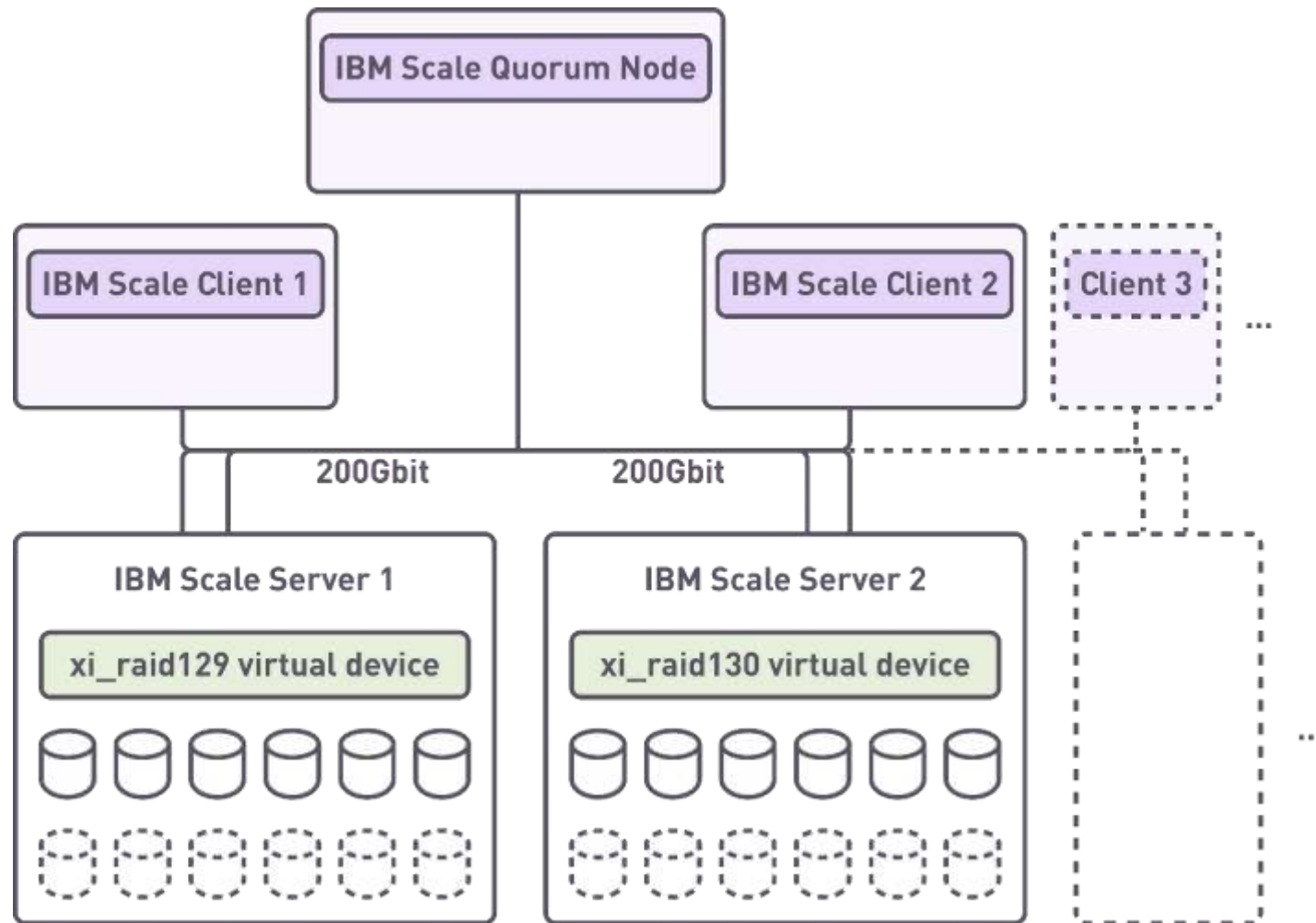
Replication set-up...

- A directory fileset 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.

...and validation

- when one of the NSD servers (NSD2) was rebooted during testing, 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

Scalability



Aalen University Case Study

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

deployed by

ABC SYSTEMS AG

APPLIED BRAINWARE & COMPUTER SYSTEMS

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

Conclusion

Combining the best of the 3 companies

MINNOR

Data protection at max speed

IBM

The best file system for HPC and AI



Flexibility and cost-optimization

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

