OSNexus | Xinnor

Dual-Layer Durability:

Scale-Up + Scale-Out EC for Cloud-Grade Storage

How public clouds achieve 11 nines of durability
— and how on-premises deployments can match it

Featuring: Xinnor xiRAID + Ceph on QuantaStor 6.7

The Single-Layer Trap

Why replica=3 alone is not enough for mission-critical workloads

Steady State

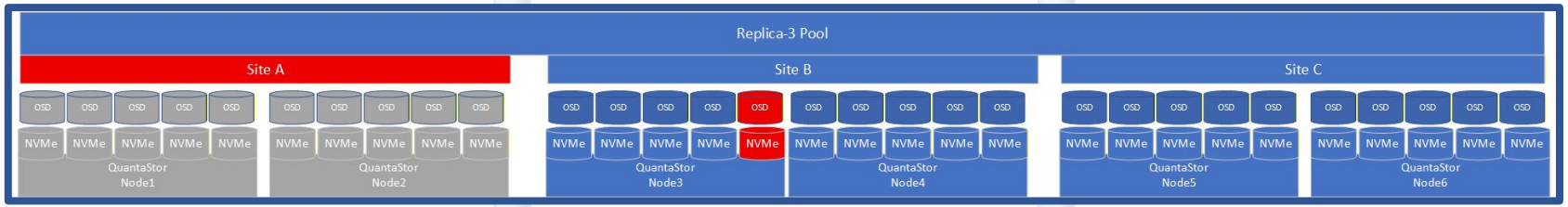
Three sites healthy. Replica=3 across sites. All looks fine.

Site Outage

Effective replica drops to 2. Cluster stays up — but you're now one failure away from crisis.

Site Outage + 1 Drive

A single OSD fails on a surviving node. Some PGs drop below min_size=2. I/O suspends. Outage.



Key insight: Drive failures are random and indifferent to your incident timeline. Designing for a single site loss is not the same as designing for a site loss ***plus*** continued hardware degradation.

The Solution: Two Independent Layers of Durability

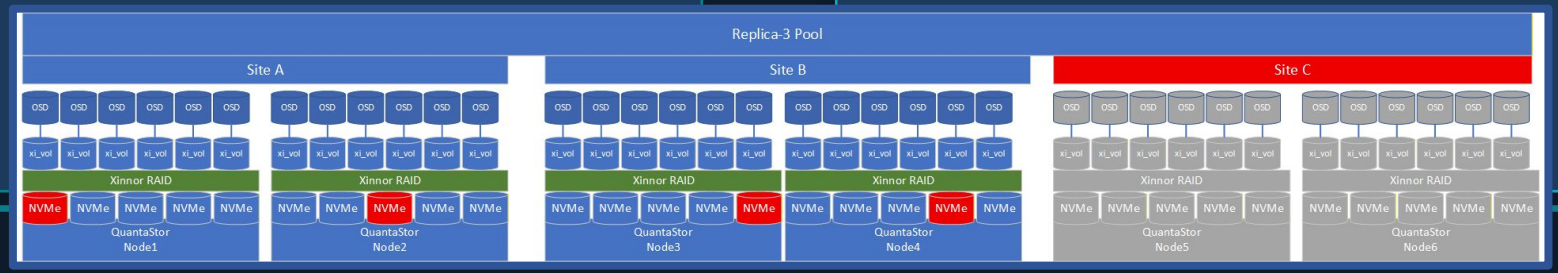
Local EC absorbs drive failures. Distributed EC absorbs node and site failures. Neither alone is sufficient.

LAYER 1 — Scale-Up / Local RAID

- Hardware or software RAID within a node
- Absorbs 2–3 simultaneous drive failures per node
- OSD stays fully online — Ceph never sees the failure
- Examples: Xinnor xiRAID, Seagate Exos Protect ADAPT
- Operates below the distributed storage layer — zero protocol changes required

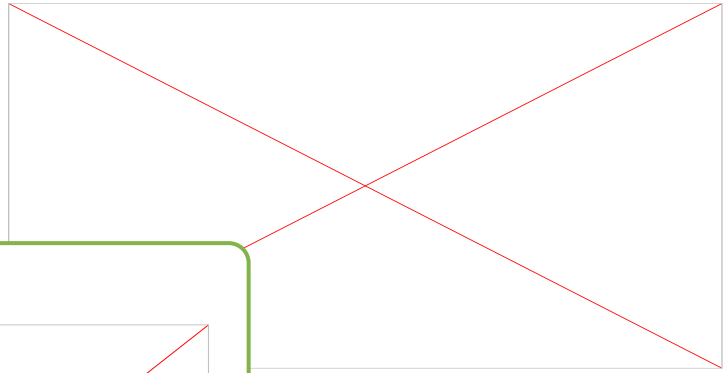
LAYER 2 — Scale-Out / Distributed EC

- Reed-Solomon EC or replication across nodes, racks, and sites
- Absorbs node, rack, or entire site failures
- Examples: Ceph replica=3, EC 4+2, 8+3, 5k+7m stretch profiles
- Operates above the local EC layer — sees only healthy logical devices
- CRUSH rules enforce failure domain separation across the cluster

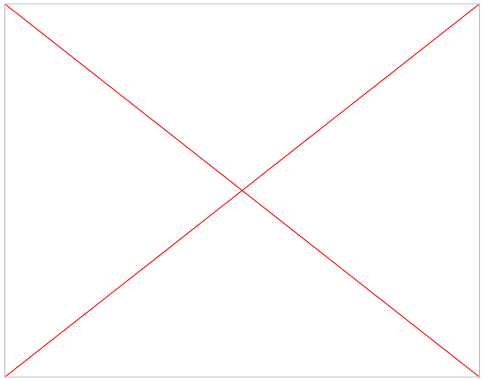


Limits of available RAID solutions

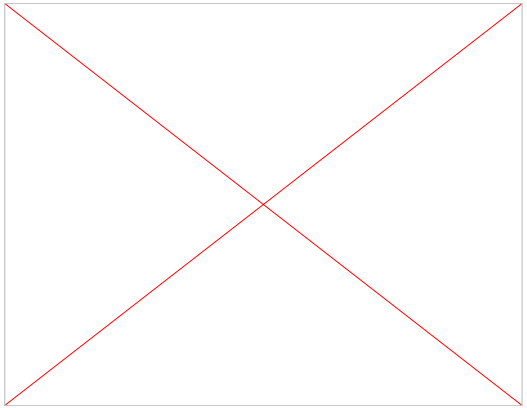
Hardware RAID card



GPU RAID



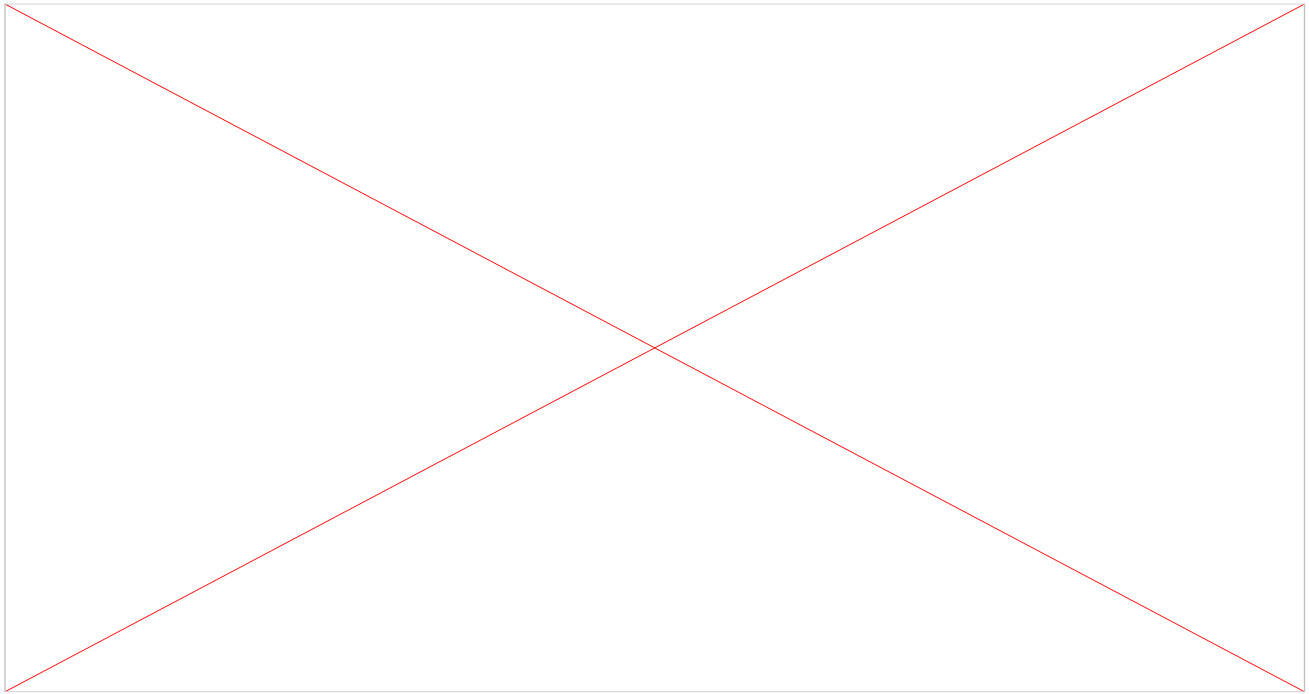
Existing software solution



Xinnor’s xiRAID Classic

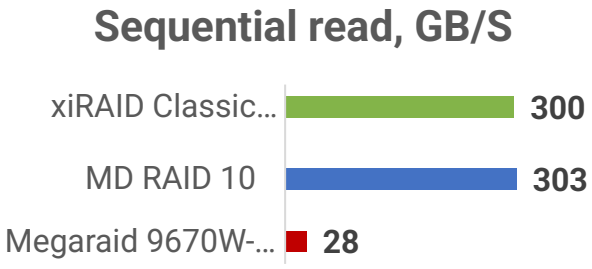
CPU-assisted RAID (AVX)

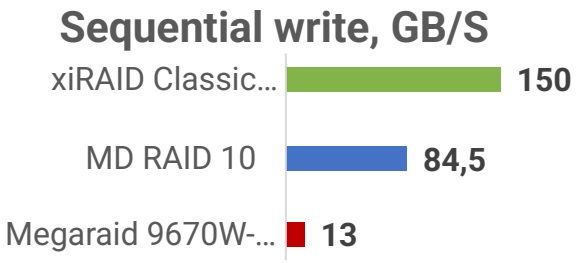
Lockless data path

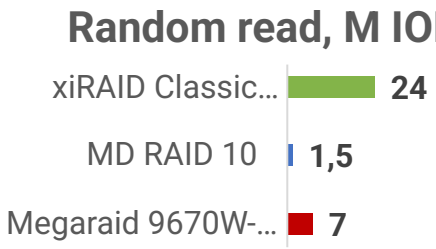


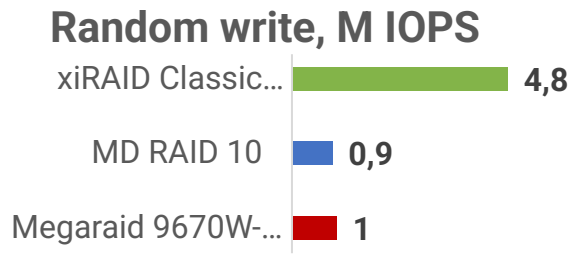
Benchmark

xiRAID and MDRAID over 24 drives while MegaRAID over 16 drives









Sources:

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

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

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

xiRAID Powers Ceph with the Same High-Performance Acceleration as Lustre

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.*

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

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

IO500

Production ISC25 List

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

#	BOF	INSTITUTION	INFORMATION: SYSTEM	INFORMATION: STORAGE VENDOR	INFORMATION:		SCORE
1	SC23	Argonne National Laboratory	Aurora	Intel			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

How Public Clouds Achieve 11+ Nines of Durability

Every major cloud uses dual-layer EC. Scale-up local parity + scale-out distributed EC.

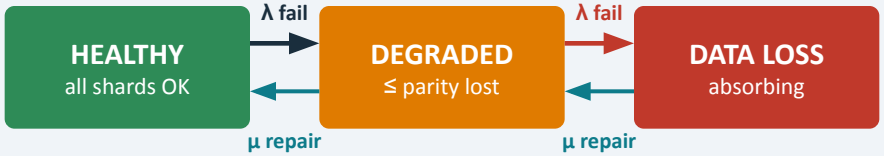
Cloud / System	Layer 1 — Local EC	Layer 2 — Distributed EC	Scheme
Google Colossus / GCS	RS(6,3) or RS(8,4) within a D-group	Cross-zone EC or replication across 3+ AZs	Reed-Solomon
Microsoft Azure LRC	LRC local parities (6,2,2) within group	LRC global parities + zone replication	Local Reconstruction Code (LRC)
Amazon S3 Standard	RS(~6,3) within an AZ (proprietary)	Cross-AZ EC or replication across 3 AZs — 11 nines	Reed-Solomon (undisclosed)
Meta f4 (OSDI 2014)	RS(10,4) within a cell / ~14 racks	RS(10,4) across cells and datacenters	Reed-Solomon (best published ref)
Backblaze Vault	RAID-6 within each Storage Pod	RS(17,3) across 20 Pods per Vault	Reed-Solomon
Wasabi Hot Cloud (WTV)	ZFS RAIDZ2/3 per storage node	RS(~14,4) across nodes and facilities	ZFS + Reed-Solomon

** Amazon S3 EC parameters are not publicly disclosed. Figures inferred from published durability claims and academic references.*

How We Prove 11+ Nines of Durability

A Markov model of each layer's failure/repair cycle yields its durability — Xinnor xiRAID and Ceph multiply to 11+ nines.

1. MARKOV STATE MODEL (per layer)



Solve for steady state

$$\text{Durability} = 1 - \pi_{\text{loss}}$$

Repair rate μ ≫ failure rate λ. xiRAID rebuilds a 15.36 TB NVMe in ~2 h, so the repair window is tiny → many nines per layer.

2. TWO INDEPENDENT LAYERS MULTIPLY

LAYER 1
Xinnor xiRAID — RAID-6, NVMe

~6 nines

15.36 TB rebuild in ~2 h

×

LAYER 2
Ceph — EC 4k+2m erasure

~6 nines

2 parity of 6 chunks, multi-node

Combined loss probability multiplies

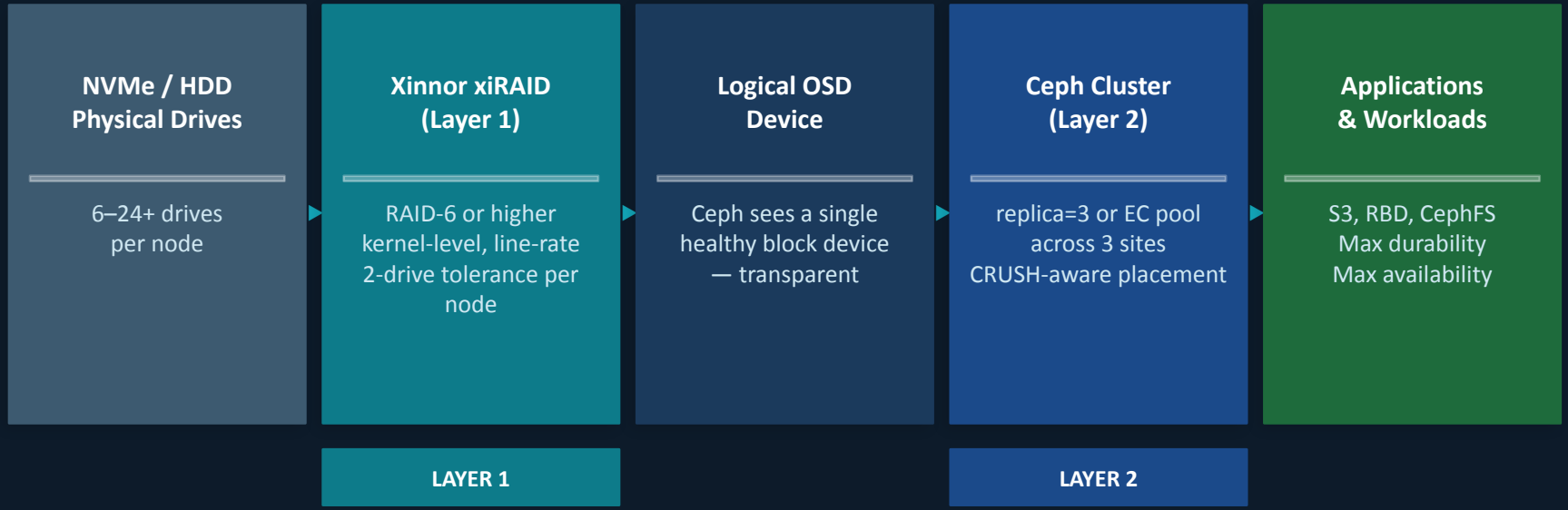
P ≈ 1e-6 × 1e-6 = 1e-12 / yr

> 11 NINES durability

Both layers tolerate two concurrent failures (RAID-6 double parity; Ceph m=2). Their independent annual loss probabilities multiply rather than add, so two ~6-nine layers compound past 11 nines. xiRAID's 2-hour NVMe rebuild keeps the per-layer repair window — and thus loss probability — extremely small.

Ceph + Xinnor xiRAID: The On-Premises Answer

xiRAID-backed OSDs bring the same dual-layer durability model to on-premises Ceph clusters.



Xinnor Integrated: Create, monitor, and manage xiRAID groups directly from the QuantaStor WUI, CLI, or API.

Compound Failure Protection: Standard OSDs vs. xiRAID-Backed OSDs

Failure Scenario	Standard OSDs	xiRAID-Backed OSDs
All three sites healthy	Full redundancy	Full redundancy
One site offline (effective replica=2)	One disk failure away from I/O suspension on affected PGs	xiRAID absorbs disk failures; OSD stays online and healthy
Site offline + 1 disk failure on a surviving node	PGs drop below min_size=2 I/O suspends — availability crisis	Cluster stays fully operational; no impact to Ceph layer
Site offline + 2 disk failures on surviving nodes	Data at risk; possible data loss depending on which PGs	Still protected (RAID-6 or higher deals with 2 drive failures)
Site offline + drive failure + recovery window (hours/days)	Extended exposure with no redundancy on some data	Continuous protection throughout the entire recovery window

xiRAID raises the floor on survivability across every failure scenario — not just the clean single-failure cases.

Durability Without the Cost Penalty

Dual-layer EC is more storage-efficient than pure replication — and achieves higher durability.

Ceph replica pool (r=3)	xiRAID RAID-6 + Ceph replica pool (r=3)	xiRAID RAID-6 + Ceph EC pool (4+2)
Raw Overhead	Raw Overhead	Raw Overhead
3.0×	**3.0×**	**1.5×**
Survives:	**Survives:**	**Survives:**
Site loss + 0 drive failures before risk	Site loss + 2 drive failures per node before risk	Node failures + 2 drive failures per node before risk
Baseline — commonly deployed but leaves a compound failure gap	*Same raw capacity overhead, dramatically higher effective durability*	*Best efficiency — cloud-grade durability at half the replica overhead*

Key Takeaways

1. **Dual-layer EC is the industry standard — not a luxury.**

   Every hyperscaler achieving 11 nines uses local EC composed with distributed EC. Single-layer replication alone leaves a compound failure gap.

2. **The compound failure scenario is not a corner case.**

   A site outage followed by a single drive failure on a surviving node is an expected, statistically likely event at scale. Design for it explicitly.

3. **xiRAID makes Layer 1 transparent to Ceph.**

   OSDs backed by xiRAID RAID-6 absorb multiple drive failures with zero impact to the Ceph cluster — no CRUSH changes, no tuning, no downtime.

4. **QuantaStor operationalizes the entire stack.**

   Native xiRAID management, monitoring, and alerting built into the same platform managing your Ceph cluster. No separate tooling.

Contact: info@osnexus.com | osnexus.com