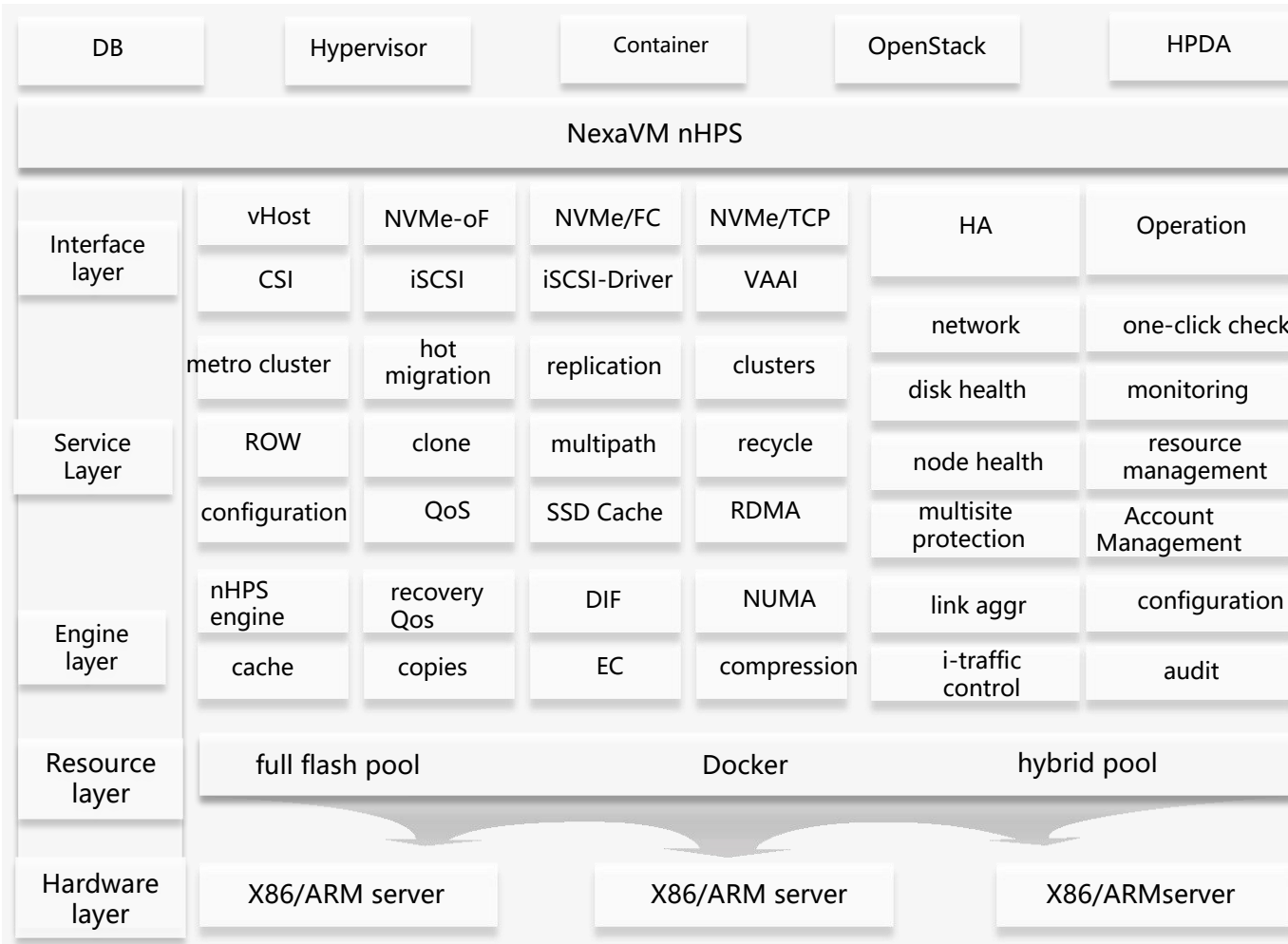


nHPS

High-Performance Distributed Block Storage

Product Architecture

Five-layer software stack with built-in reliability and operations planes



Design Principles

-  **IP-based fabric.** No special network devices required — fits the vast majority of application scenarios.
-  **No hardware lock-in.** Runs on x86 and domestic ARM servers; new & old hardware share one pool, multi-vendor hardware in one cluster — flexible TCO control.
-  **Mixed media.** SSD, SAS and SATA combined freely to match the performance profile of each workload.
-  **Generic hardware acceleration.** Fully exploits the potential of commodity processors and disks.
-  **Cloud-native delivery.** Kubernetes-orchestrated deployment of block storage services on any cloud or physical environment.

| Feature Highlights

Ten engineering choices that define nHPS

10

capability pillars
on one storage platform



Proprietary I/O Stack

RTC run-to-completion,
fully lock-free end to end



NUMA-Aware

End-to-end zero copy,
huge-page allocator



Dual-Stack Network

TCP + RoCE v2, chosen
per storage pool



Snapshot / Clone

Lossless ROW, up to 10K
snapshots per volume



All-Flash / Hybrid

NVMe all-flash or SSD-
cache + HDD pools



Cloud-Native CSI

Native Kubernetes
volume provisioning



Sync / Async Replication

Metro active-active and
remote protection



Online Migration

Cross-pool, cross-cluster,
business unseen



Reliability Portfolio

Erasure coding, replicas, multi-
level fault domains

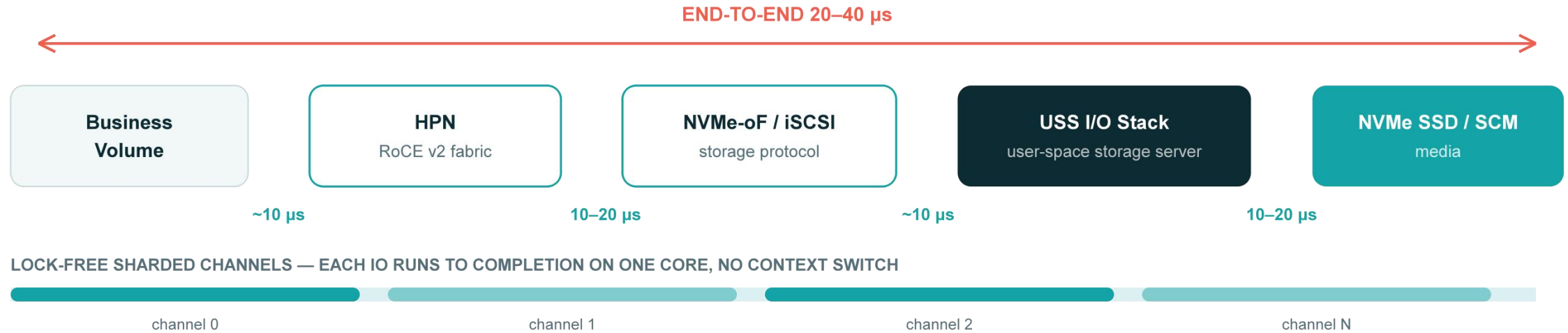


Multi-Platform

VMware, OpenStack, K8s
and more

Proprietary High-Performance I/O Stack

End-to-end ultra-low latency on a fully lock-free, run-to-completion architecture



20–40 μ s

end-to-end IO latency

Measured on the self-developed ultra-speed stack



Run-to-completion, fully lock-free. IOs are sharded by ID into independent channels and driven end-to-end to disk; resources are partitioned across the whole path — zero locks anywhere.



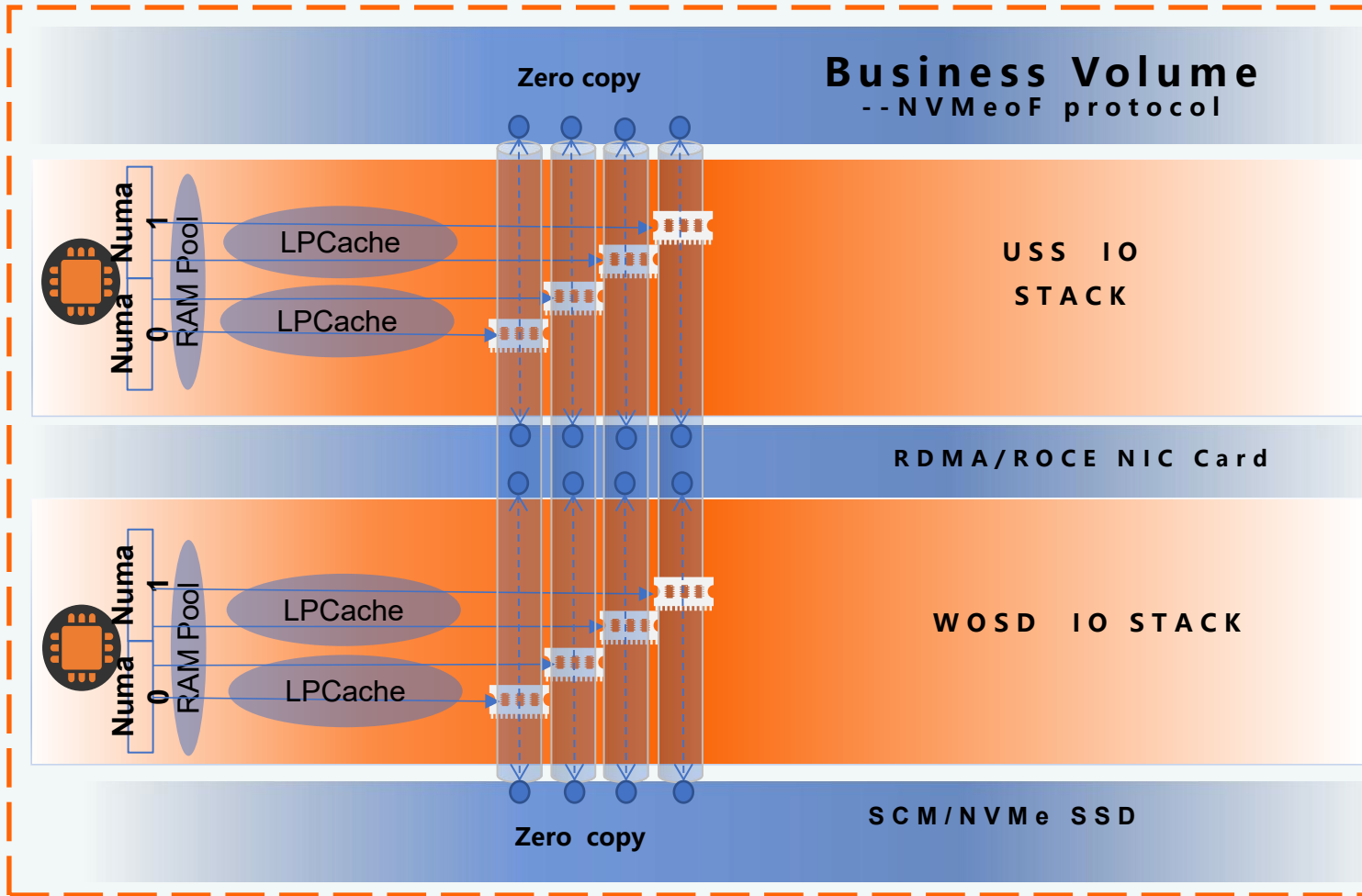
Unlocks Optane-class media. The 20–40 μ s stack fully exploits Intel Optane SSDs and outperforms conventional NVMe SSD deployments by a wide margin.



Minimal path, extreme experience. The leanest possible IO software stack is what gives the ultra-speed storage pool its performance edge.

NUMA-Aware Framework — End-to-End Zero Copy

Data never leaves its NUMA domain from NIC to disk



End to end Zero copy

Compute-side zero-copy

Implement zero-copy across the entire stack—from protocol-side reception, through I/O request buffer allocation, to network transmission.

Storage-side zero-copy

Implement zero-copy across the entire stack

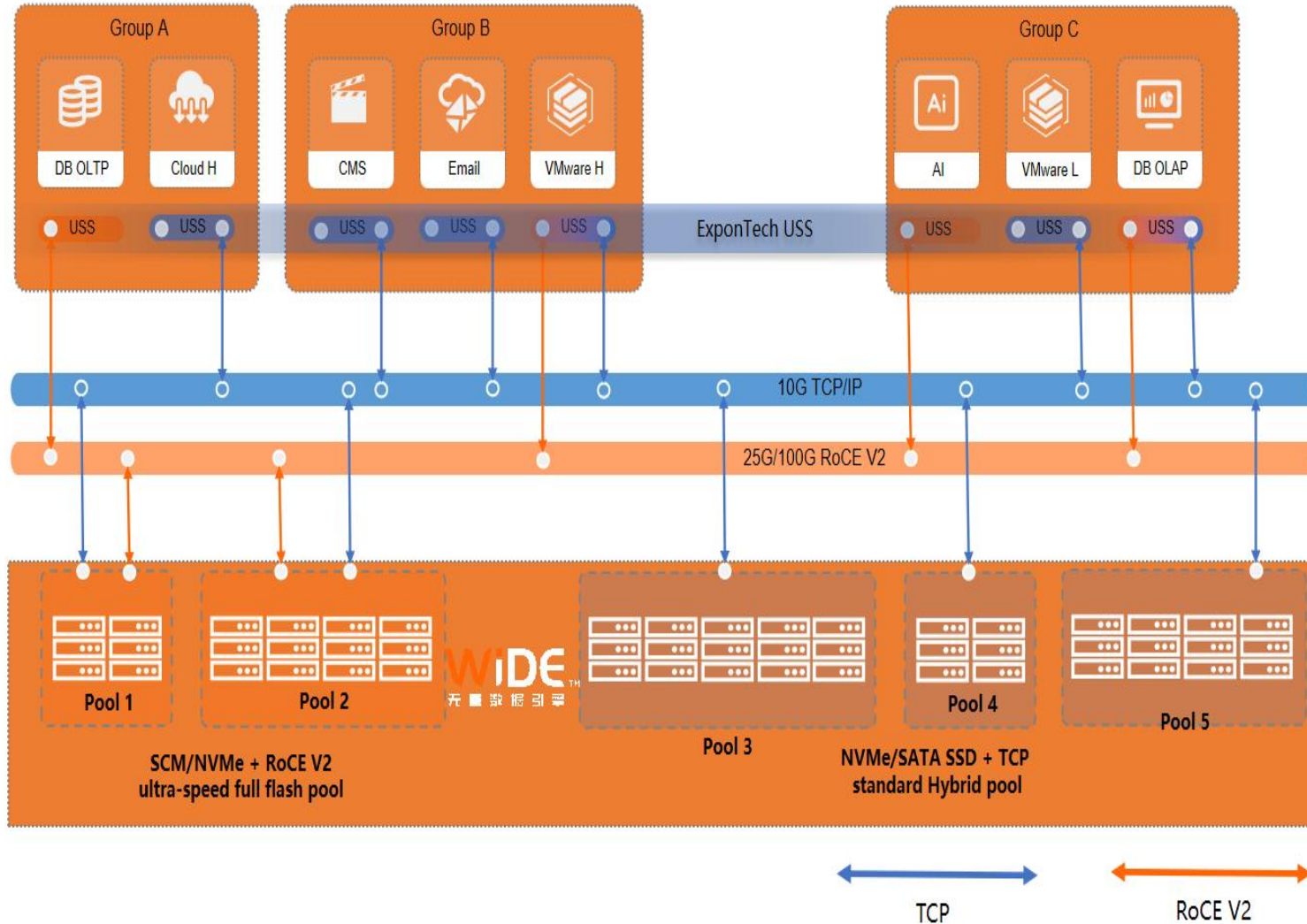
Huge-page memory management

Centrally manage huge-page memory and implement a dedicated memory allocator.

Allocate memory within the same NUMA node whenever possible to improve memory locality and reduce cross-NUMA access overhead.

Proprietary Dual-Stack Network

Extract the full potential of every link speed, with highly reliable interconnect



- Support thread-level dual-stack networking with ****TCP and RoCEv2****.
- Different workloads can use different network protocols based on price/performance targets.

High-Performance Erasure Coding

N data + M parity stripes — any M failures, zero data loss

ONE EC STRIPE ACROSS THE CLUSTER (4+2 EXAMPLE)



D = data block · C = parity block — encoded in the engine, compressed and DIF-protected in software

Any 2 of the 6 blocks can be lost — rebuilds start in milliseconds

Rack-level and node-level redundancy policies, configurable per pool

91.6%
maximum space utilization
EC ratios from 2+1 up to 22+2,
balanced against 3-replica overhead

100 μ s-class response

With the near-replica pool, the system answers in $\sim 100 \mu$ s inside its normal operating band (40–60% load).

Millisecond failover

Thanks to the Append-only engine, sub-healthy disks or links switch over in milliseconds — business stays stable.

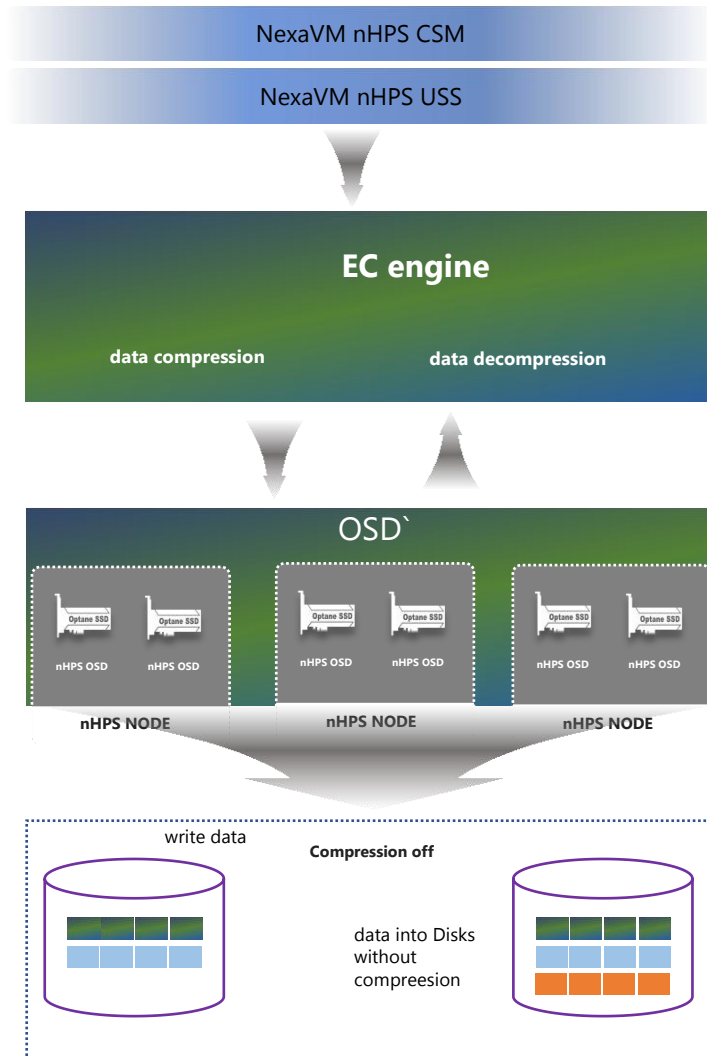
Longer SSD life

Append-only writes avoid overwrite and random-write churn, sharply reducing flash erase cycles and extending SSD lifespan.

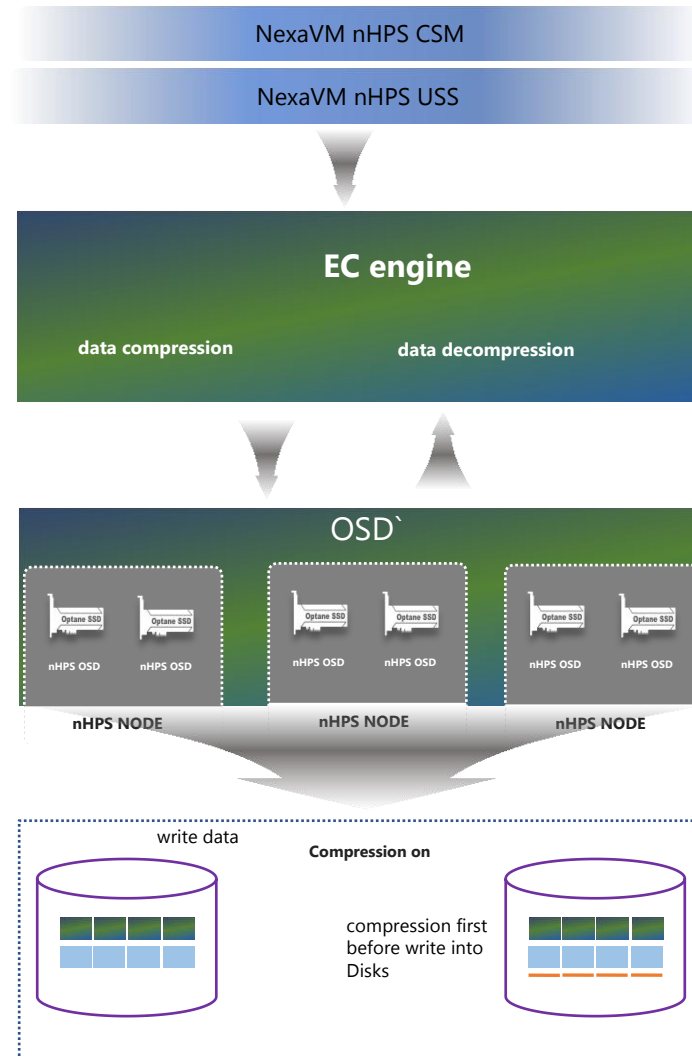
Dynamic EC: the EC ratio can be tuned per volume — or adjusted automatically as hardware scales — so capacity and cost stay in balance at every stage.

Compression -in roadmap

Compression off



Compression on



>2:1

Compression Ratio

100%

data completeness

ROW Snapshot / Clone

Redirect-on-write — high performance snapshots and instant clones

1 · LUN BEFORE SNAPSHOT



snapshot
→

2 · SNAPSHOT 0 (FROZEN VIEW)



writes
→

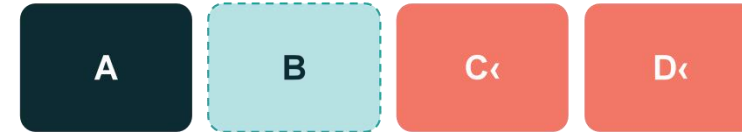
Write B,C
new blocks B< C<
originals kept
snapshot untouched

3 · LIVE LUN AFTER WRITES — REDIRECT-ON-WRITE



Only changed blocks are redirected — the snapshot never degrades, generation after generation.

4 · CLONE LUN — WRITE C,D



Clone shares unmodified blocks instantly — no full copy, no capacity spike.

Why redirect-on-write beats copy-on-write:

no read-modify-write penalty on the live volume, snapshot chains stay at first-generation quality forever.

Source deck: "truly lossless snapshot" — data written after the snapshot never touches the snapshot itself.

0 loss

True lossless snapshots — originals never overwritten

Seconds

Second-level snapshot granularity

10K+

Snapshots per single volume

Multi-policy

Multiple backup policies per volume

All-Flash / Hybrid & HCI/Storage

All-Flash Form

SDS 2.0

Pure NVMe SSD + SATA SSD media — for databases, AI and latency-critical tiers

NVMe SSD

SATA SSD

RoCE v2 ready

Network options — on demand

100GE / 200GE IB

100GE RDMA

25GE RDMA

10GE TCP

Protocols: Vhost · NVMe-oF · iSCSI over RDMA/TCP — every pool exposes the interface its consumers need.

Hybrid-Flash Form

SDS 2.0

SSD cache + HDD combination — cost-optimized capacity for general workloads

SSD Cache

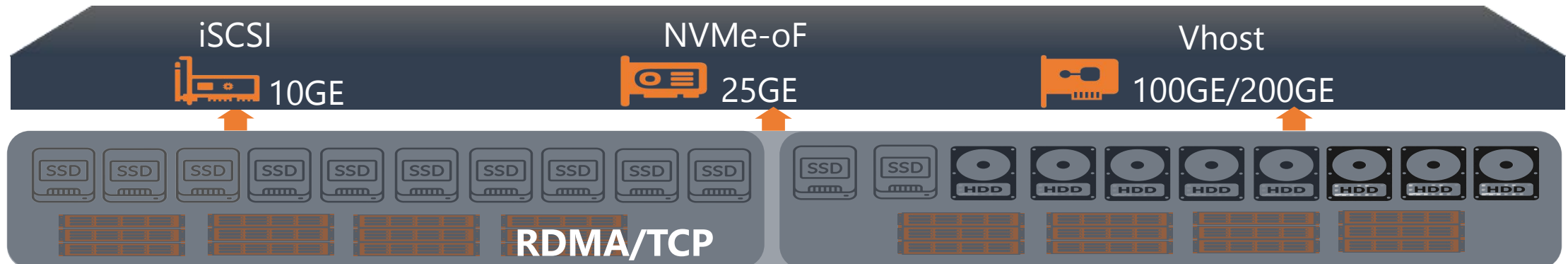
HDD

Converged deployment

Fully containerized install; hyper-converged form factor supported.

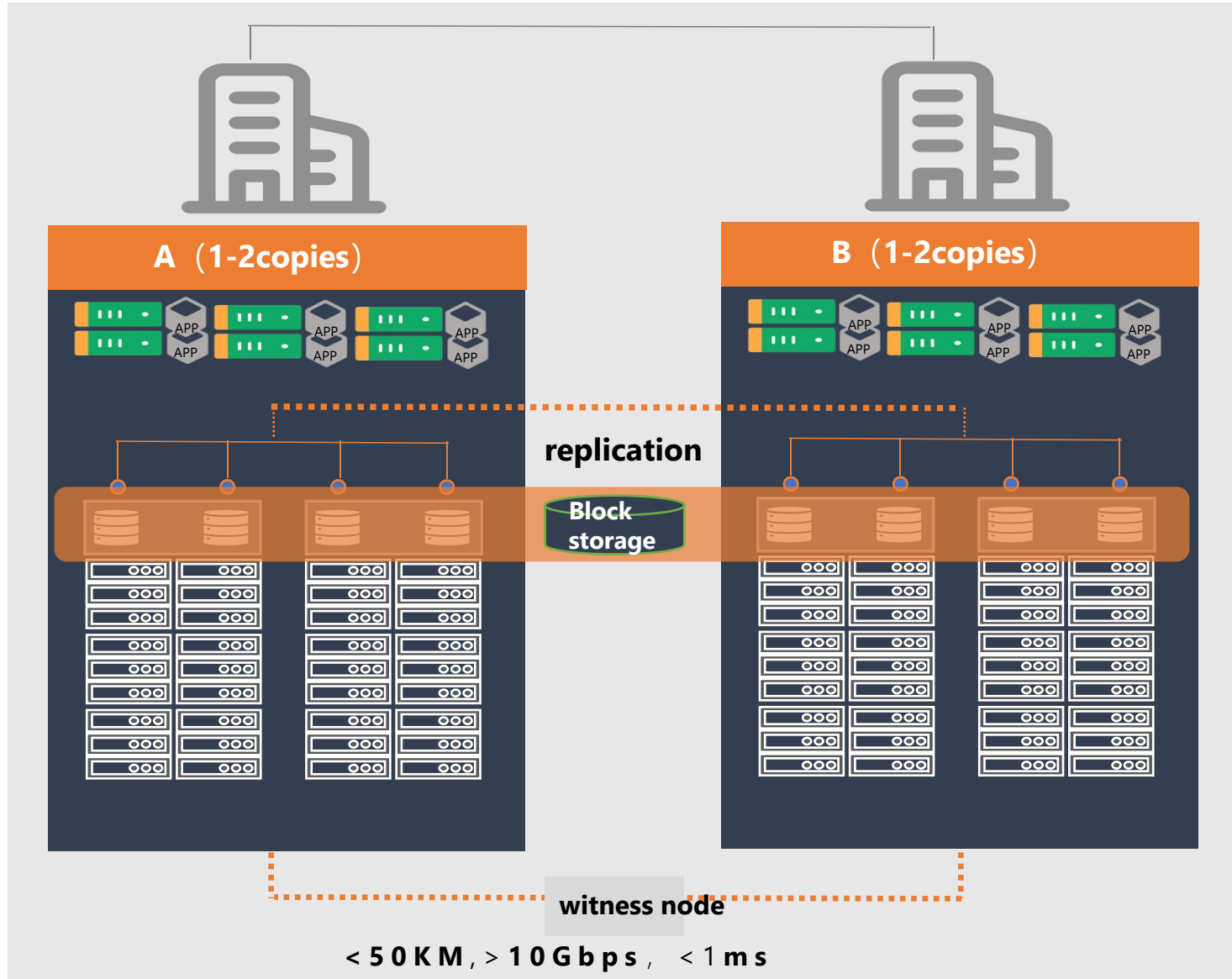
Disaggregated deployment

Dedicated storage cluster serving multiple compute platforms.



Stretched Cluster

Synchronous replicas across two data centers in the same city



Protection: metro active/active across two campuses — multi-replica data written to both data centers in real time via stretched-cluster cross-copy.

zero data loss: RPO=0, RTO = switchover time.

Safe Migration Within and Across Clusters

Online, log-based volume migration — business completely unaware

Log-based mechanism

Consistency is guaranteed at IO granularity throughout migration, achieving second-level RPO and RTO. Snapshot-based alternatives are capped at minutes to hours by snapshot create/delete and merge performance.

Online hot migration, volume granularity

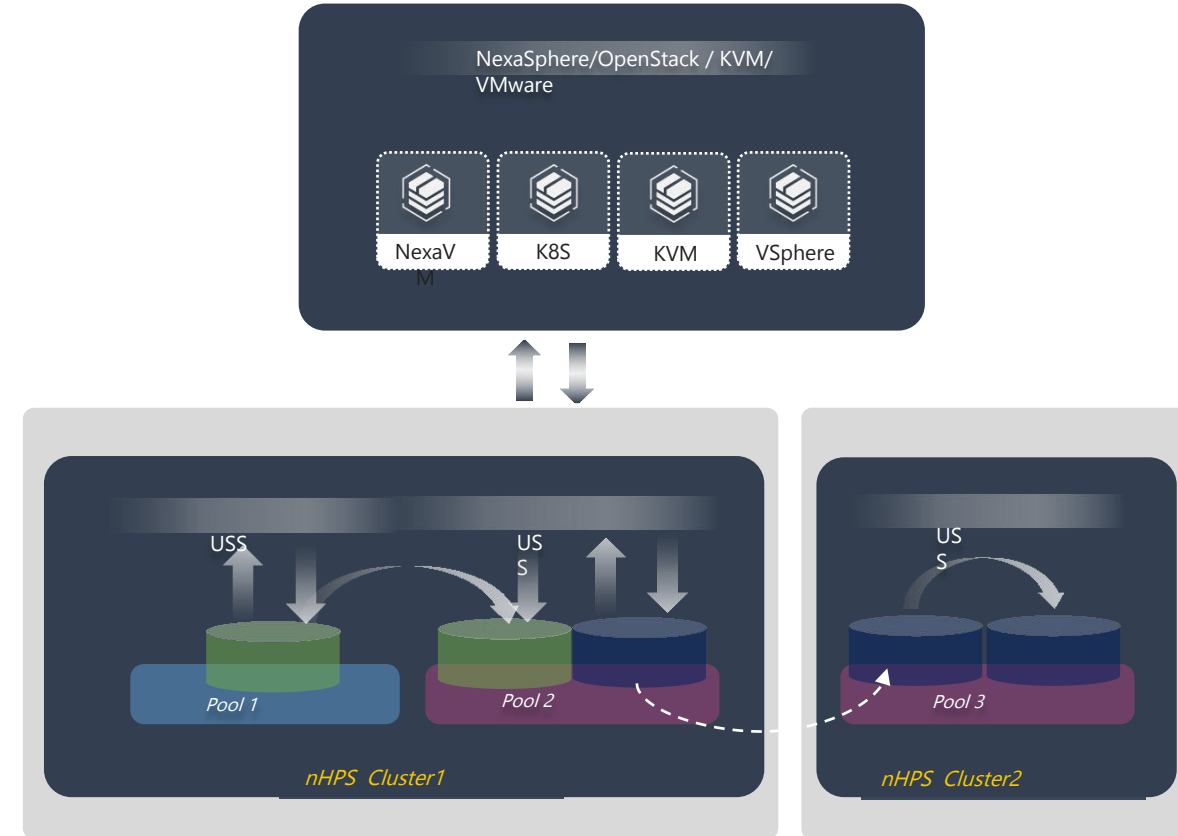
Volumes move across pools within a cluster — or across clusters — while business keeps running. No coordination with upper-layer applications required.

Balance capacity and performance

Use migration to equalize capacity across pools and clusters, to spread load, and to move data between performance tiers as application requirements change.

nHPS vs Ceph

Unlimited migration count (Ceph's design restricts it), parallel replication for faster data movement, and full visualization of progress.



Multi-Dimensional Reliability & Availability



Replica consistency check

Three replicas are verified and repaired periodically or on fault triggers — eliminating the risk of silent disk errors before they surface.



Network sub-health detection

Monitors latency anomalies, error packets and packet loss across links, ports and NICs — degradation is caught long before failure.



Disk sub-health detection

Intelligent algorithms analyze disk faults, slow disks and remaining life to act on risk before it becomes an outage.



Multi-level fault domains

Disk-level, server-level and rack-level fault domains isolate blast radius with far more flexibility than flat clusters.



Data flow across physical pools

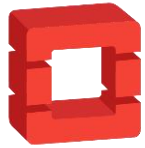
Data managed across multiple pools spreads fault domains apart and lets the strengths of each pool — performance or capacity — be exploited.



Active-active DR

Strongly consistent replica synchronization between two regions keeps business continuity fully assured — RPO=0 for metro distances.

Multi-Platform Support



OpenStack

iSCSI



VMware

iSCSI



iSCSI/vHost



K8S

CSI



iSCSI

OS



processor



Performance Unlike Anything Before

Measured on commodity x86 servers · 200G RoCE v2 · high-performance NVMe SSD

118M

IOPS · 4K random read

32 nodes · 0.368 ms latency

And 55M IOPS random write at 0.364 ms — the performance bottleneck is simply the hardware you buy.

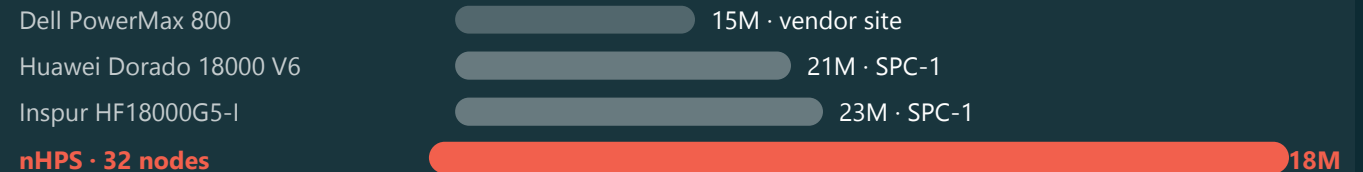
3 insights from the scaling curve:

- Performance grows linearly with nodes
- Latency stays almost flat as the cluster grows
- Scale out = scale up, with zero re-architecture

LINEAR SCALING — MEASURED, NOT PROJECTED

Nodes	4K random read	4K random write
3	11.7M IOPS · 0.35 ms	5.9M IOPS · 0.36 ms
10	39M IOPS · 0.319 ms	18.5M IOPS · 0.356 ms
20	71M IOPS · 0.375 ms	36M IOPS · 0.37 ms
32	118M IOPS · 0.368 ms	55M IOPS · 0.364 ms

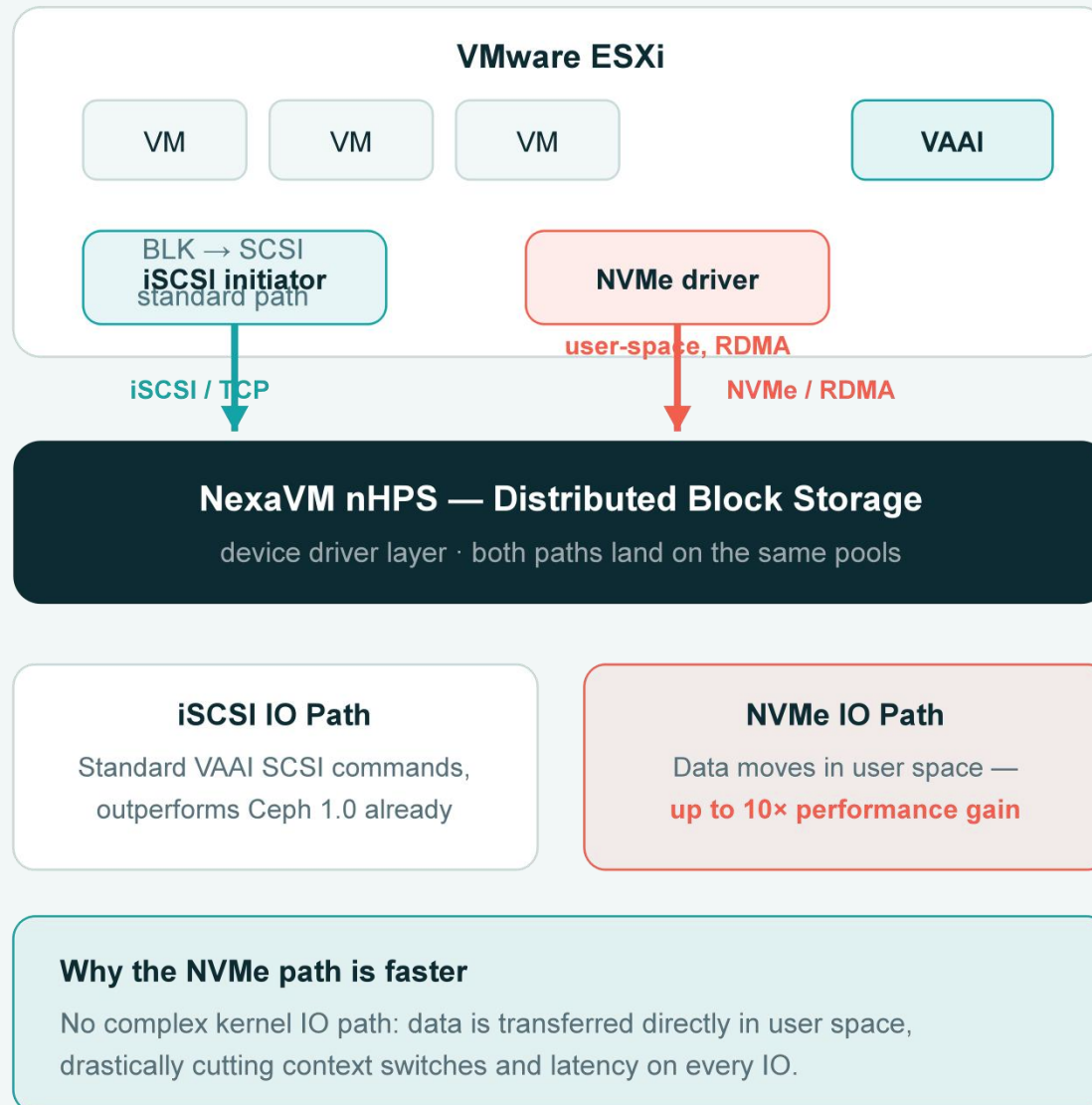
VS 32-CONTROLLER HIGH-END ARRAYS (PUBLIC SPECS)



Bar length proportional to published IOPS. Array figures from vendor sites and spcreports.org.

nHPS — VMware Solution

Native VAAI integration with two IO paths — iSCSI and NVMe/RDMA



Native vSphere integration

Standard VMware VAAI command support makes nHPS a first-class vSphere datastore

Two protocols, one platform

Choose iSCSI or RDMA (NVMe-oF) per business need — network choice follows application performance requirements.

10x

performance uplift on the NVMe path versus traditional complex IO stacks — while even the iSCSI path beats Ceph 1.0.

Thank You

nHPS — High-Performance Distributed Block Storage

NexaVM · Technical Presentation