croit

BENCHMARK

croit FlashStore OSD: more performance on the same hardware

croit FlashStore OSD against BlueStore on NVMe clusters (30 OSDs; 5 and 10 OSDs for 4 MiB objects): more throughput, lower latency, less CPU per operation and modest memory use, measured with unmodified Ceph clients (stock librados) against Ceph 21.2.0 BlueStore.

croit FlashStore OSD is available as a preview for testing and evaluation. It is not yet recommended for production data. croit FlashStore OSD is included in the croit frontier release.



Keep it yours.

Storage, virtualization and AI, all in house.

Version 09 Nov 2026 17:01

croit

The same hardware, a faster storage engine

Two NVMe clusters (30 OSDs; 5 and 10 OSDs for 4 MiB), 3 clients, identical cells and data volumes for both OSD engines.

+50 %

Throughput, replicated (size 3)

up to +123 % (4K seq read)

Average over 10 rep3 cells on the 30-OSD cluster (64K / 4K, read, write, mixed), 48 clients x QD16, IOPS and MB/s

+54 %

Throughput, erasure coded 2+2

up to +187 % (4K rand write)

Average over 10 EC 2+2 cells on the 30-OSD cluster, same cell set and load

-52 %

p99 latency, replicated

up to -94 % (4K rand read)

Average over 10 rep3 cells, 30 OSDs (p99 client completion latency, lower is better)

-36 %

OSD CPU per operation

up to -59 % (4K seq write)

Average over 10 rep3 cells, OSD CPU microseconds per op

2-5 GiB

Memory per OSD (median at cell end)

FlashStore OSD (process + hugepages): 2.2-4.8 GiB (median per OSD at cell end), kernel cache at most 0.4 GiB

Median over 30 OSDs, 28 cells. BlueStore: process 2.9-4.0 GiB plus kernel block cache charged to the OSD, up to 26 GiB in the EC 4K cells

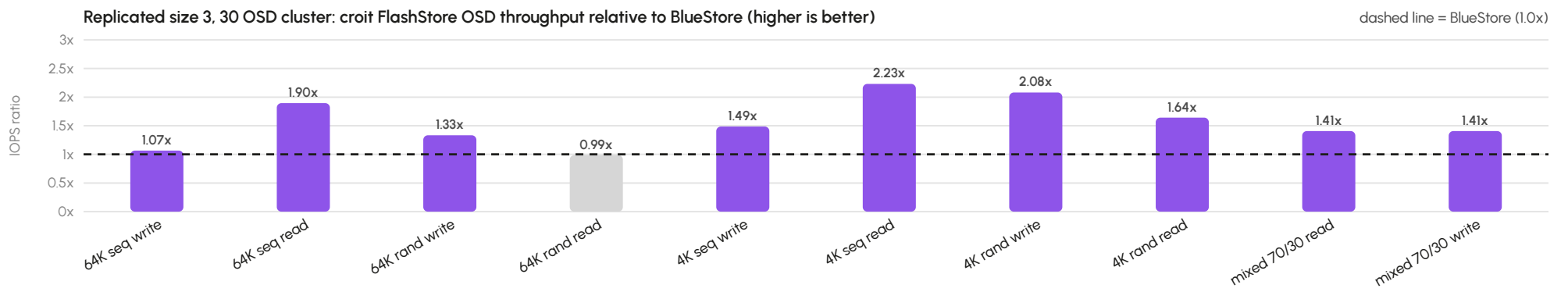
Also: replicated IOPS at queue depth 1 (QD1): +9 % on average over 8 cells. Large 4 MiB objects (pages 5 and 6), throughput averaged over 7 load points each: 5 OSDs: random write +14 %, random read +1 %; 10 OSDs: random write +28 %, random read +4 %; write p99 at 1,536 in flight 5 OSDs -69 %, 10 OSDs -45 %. Raw capacity use after the same test cycle is on par (page 12). Average = geometric mean of the per-cell ratio croit FlashStore OSD / BlueStore over the stated cell family; "up to" = best single cell. All numbers: kernel TCP, no RDMA; measurement conditions on pages 7 and 8.

© 2026 croit GmbH

croit FlashStore OSD is available as a preview for testing and evaluation. It is not yet recommended for production data.

Up to 123 % more throughput, replicated

Replicated throughput per cell on the 30-OSD cluster; 48 clients x QD16 (768 I/Os in flight).



Cell	BlueStore IOPS	FlashStore IOPS	Change
64K seq write	375,783	400,838	**+7 %**
64K seq read	522,622	990,806	**+90 %**
64K rand write	329,119	438,560	**+33 %**
64K rand read	1,060,164	1,045,845	-1 %
4K seq write	612,167	909,481	**+49 %**
4K seq read	989,827	2,209,461	**+123 %**
4K rand write	470,596	979,347	**+108 %**
4K rand read	1,450,914	2,380,557	**+64 %**
mixed 70/30 read	334,981	471,237	**+41 %**
mixed 70/30 write	143,576	201,957	**+41 %**

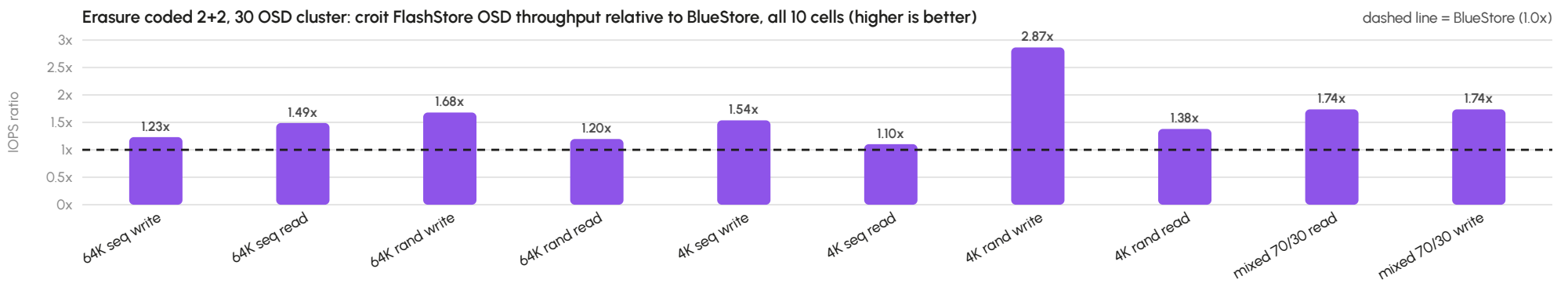
30 OSD cluster (64K, 4K, mixed cells)

30 OSDs (5 hosts x 6), 6 cores each, 1 NVMe each
Network: bond0 2 x 200 Gbit/s (LACP), MTU 1500
Transport: kernel TCP, no RDMA (details page 7)

Average +50 % (geometric mean of 10 cells). Bars = IOPS ratio; purple = FlashStore at or above BlueStore, grey = below. 4 MiB objects: pages 5 and 6.

Erasure coded 2+2: +54 % on average

Erasure coded throughput per cell on the 30-OSD cluster; 48 clients x QD16.



Cell	BlueStore IOPS	FlashStore IOPS	Change
64K seq write	259,503	319,737	+23 %
64K seq read	371,786	552,945	+49 %
64K rand write	147,056	246,991	+68 %
64K rand read	594,239	711,650	+20 %
4K seq write	324,304	498,078	+54 %
4K seq read	1,043,835	1,150,923	+10 %
4K rand write	141,414	405,178	+187 %
4K rand read	1,378,662	1,901,424	+38 %
mixed 70/30 read	164,094	285,006	+74 %
mixed 70/30 write	70,275	122,161	+74 %

30 OSD cluster (64K, 4K, mixed cells)

30 OSDs (5 hosts x 6), 6 cores each, 1 NVMe each
Network: bond0 2 x 200 Gbit/s (LACP), MTU 1500
Transport: kernel TCP, no RDMA (details page 7)

Average +54 % over the 10 EC 2+2 cells (geometric mean), up to +187 % (4K rand write). 4 MiB EC cells are not shown: on 30 OSDs they were bound by the client network (69 to 73 GB/s read on both engines); no 5- or 10-OSD EC run exists. Bars = IOPS ratio; purple = FlashStore at or above BlueStore.

croit

4 MiB objects on 5 OSDs: write +14 %, read +1 % on average

5 NVMe OSDs (one per host, 6 cores each), replicated size 3, 128 PGs, 512 GiB data set, 3 client hosts; caches dropped before every cell.

+14 %

Random write throughput

up to +26 % (192 in flight)

Geometric mean over 7 load points; sequential write (fill): +7 % (9.59 to 10.2 GB/s)

-69 %

Write p99 at 1,536 in flight

8.9 s to 2.7 s

Lower at 7 of 7 write load points; geometric mean over the 7 points -46 %, best -69 % (1,536 in flight)

+1 %

Random read throughput

up to +29 % (1 stream, QD 1)

Geometric mean over 7 load points: 1 % above BlueStore on average; at or above BlueStore at 3 of 7 points

+9 %

Read p99 latency

best -20 % (1 stream, QD 1)

Geometric mean over 7 load points; lower at 2 of 7 points; at 1,536 in flight: 3.0 s to 3.4 s

Write

Load point	BlueStore GB/s	FlashStore GB/s	Change	p99 BlueStore ms	p99 FlashStore ms
Sequential write (fill, 96 in flight)	9.59	10.2	**+7 %**	283	**73.9**
1 stream, QD 1	1.28	1.53	**+20 %**	3.9	**2.8**
1 stream, QD 4	3.95	4.79	**+21 %**	6.3	**5.4**
1 stream, QD 16	7.68	8.74	**+14 %**	21.6	**17.2**
48 in flight	9.29	9.06	-2 %	89.7	**43.3**
192 in flight	9.33	11.8	**+26 %**	541	**234**
768 in flight	10.9*	11.8	**+8 %**	2,835	**1,317**
1,536 in flight	10.6*	11.7	**+11 %**	8,925	**2,735**

Random read

Load point	BlueStore GB/s	FlashStore GB/s	Change	p99 BlueStore ms	p99 FlashStore ms
1 stream, QD 1	2.14	2.76	**+29 %**	2.4	**1.9**
1 stream, QD 4	6.50	7.06	**+9 %**	5.0	5.3
1 stream, QD 16	10.1	8.53	-16 %	24.2	29.8
48 in flight	28.6	28.8	**+1 %**	24.5	26.1
192 in flight	50.5	49.6	-2 %	192	**156**
768 in flight	50.5	49.4	-2 %	952	1,636
1,536 in flight	50.6	48.2	-5 %	2,970	3,439

Load point = ops in flight: 1 stream = 1 client host x 1 job x QD 1/4/16; N in flight = 3 client hosts x N/48 jobs x QD16. At 1,536 in flight reads run at 50.6 GB/s (BlueStore) and 48.2 GB/s (FlashStore). Why 4 MiB is shown on 5 and 10 OSDs: on 30 OSDs the 4 MiB cells were bound by the client network (69 to 73 GB/s on both engines), which says nothing about the OSD; on 5 and 10 OSDs the OSDs are the limit. Single run per cell. * Not steady within the 90 s window, window average shown: BlueStore random write 768 in flight (middle third 9.3, last third 13.6 GB/s); BlueStore random write 1,536 in flight (middle third 9.3, last third 12.6 GB/s).

croit

4 MiB objects on 10 OSDs: write +28 %, read +4 % on average

10 NVMe OSDs (two per host, 6 cores each), replicated size 3, 256 PGs, 1 TiB data set, 3 client hosts; caches dropped before every cell.

+28 %

Random write throughput

up to +36 % (192 in flight)

Geometric mean over 7 load points; sequential write (fill): +35 % (16.7 to 22.6 GB/s)

-45 %

Write p99 at 1,536 in flight

5.0 s to 2.8 s

Lower at 7 of 7 write load points; geometric mean over the 7 points -45 %, best -80 % (192 in flight)

+4 %

Random read throughput

up to +23 % (1 stream, QD 1)

Geometric mean over 7 load points: 4 % above BlueStore on average; at or above BlueStore at 3 of 7 points

+2 %

Read p99 latency

best -8 % (48 in flight)

Geometric mean over 7 load points; lower at 5 of 7 points; at 1,536 in flight: 600 ms to 591 ms

Write

Load point	BlueStore GB/s	FlashStore GB/s	Change	p99 BlueStore ms	p99 FlashStore ms
Sequential write (fill, 96 in flight)	16.7*	22.6	**+35 %**	179	**39.6**
1 stream, QD 1	1.15	1.46	**+27 %**	4.4	**2.9**
1 stream, QD 4	4.18	5.36	**+28 %**	5.7	**4.4**
1 stream, QD 16	10.6	13.0	**+23 %**	12.9	**10.0**
48 in flight	15.9	20.5	**+29 %**	39.1	**24.5**
192 in flight	17.0	23.2	**+36 %**	583	**115**
768 in flight	17.6	23.0	**+31 %**	2,298	**1,267**
1,536 in flight	18.6*	22.9	**+23 %**	5,000	**2,768**

Random read

Load point	BlueStore GB/s	FlashStore GB/s	Change	p99 BlueStore ms	p99 FlashStore ms
1 stream, QD 1	2.03	2.51	**+23 %**	2.5	**2.4**
1 stream, QD 4	6.34	6.99	**+10 %**	5.1	**5.1**
1 stream, QD 16	10.5	10.3	-2 %	29.5	31.1
48 in flight	30.0	31.9	**+6 %**	32.1	**29.5**
192 in flight	59.2	57.6	-3 %	67.6	83.4
768 in flight	73.1	72.1	-1 %	296	**292**
1,536 in flight	74.7	73.3	-2 %	600	**591**

Load point = ops in flight: 1 stream = 1 client host x 1 job x QD 1/4/16; N in flight = 3 client hosts x N/48 jobs x QD16. At 1,536 in flight reads run at 74.7 GB/s (BlueStore) and 73.3 GB/s (FlashStore). Single run per cell. * Not steady within the 90 s window, window average shown: BlueStore sequential write (fill) (middle third 17.4, last third 14.9 GB/s); BlueStore random write 1,536 in flight (middle third 16.6, last third 22.5 GB/s).

croit

How the results were measured: cluster, network, transport

Every value is read from the run data of the measurements shown on the result pages; 30-OSD cluster and 5-/10-OSD clusters side by side.

	30 OSD cluster (64K, 4K, mixed; latency, CPU, memory, capacity)	5 and 10 OSD clusters (4 MiB)
Measured	BlueStore reference 2026-11-02, FlashStore 2026-11-07; same cluster, same test specification (not interleaved)	BlueStore 2026-11-03, FlashStore 2026-11-07 (UTC); same cluster, same test specification
Cluster	8 hosts: 5 OSD hosts x 6 OSDs = 30 OSDs, 3 client hosts; 1 NVMe per OSD	same hosts; 5 OSD cluster: 5 OSDs (one per OSD host); 10 OSD cluster: 10 OSDs (two per OSD host); in their own CRUSH root, 1 NVMe per OSD; the other OSDs are not in the pool
Client network	bond0 = LACP 2 x 200 Gbit/s = 400 Gbit/s per host, MTU 1500	bond0 = LACP 2 x 200 Gbit/s = 400 Gbit/s per host, MTU 1500
Cluster network	same bond0 as the client network (public = cluster network), MTU 1500	same bond0 (public = cluster network), MTU 1500
Transport	BlueStore: stock Ceph messenger over kernel TCP. FlashStore: kernel TCP over io_uring (RDMA off, no RDMA counters). Clients: stock librados over kernel TCP	same: kernel TCP for BlueStore, FlashStore and clients; RDMA off
Kernel	7.2.13-1.el9.elrepo.x86_64 on OSD hosts and client hosts	7.2.13-1.el9.elrepo.x86_64 on OSD hosts and client hosts
Ceph / build	BlueStore: Ceph 21.2.0_frontier0 (b060042deb05); FlashStore: build 422be2684899, Ceph 21.2.0 compatible	BlueStore: Ceph 21.2.0_frontier0 (b060042deb05); FlashStore: build 422be2684899 (same as the 30 OSD cluster)
Store tier	FlashStore: kernel NVMe tier (io_uring passthrough); BlueStore: stock, on the kernel block device	same; BlueStore deployed with ceph-volume lvm on the whole drive
Hardware	AMD EPYC 9455P, 48 cores, SMT off, 503 GiB RAM per host; NVMe SAMSUNG MZ3LO15THBLA-00A07	same hosts and drive model
CPU per OSD	6 dedicated cores per OSD (one L3 domain, cpuset), both engines	5 OSD cluster: 6 dedicated cores per OSD (CPUs 30-35), both engines; 10 OSD cluster: 6 dedicated cores per OSD (CPUs 24-29, 30-35), both engines

SCOPE OF THE COMPARISON: BLUESTORE, NOT CRIMSON

Crimson/SeaStore is the upstream next-generation OSD and is still under development. Igor Fedotov, a long-time BlueStore developer, tried repeatedly to benchmark Crimson/SeaStore on the same hardware and could not get it to complete a successful benchmark run. BlueStore is what runs in production today, so it is the relevant reference.

croit

How the results were measured: load, data, configuration

Same cells, same load definition and same settle rule for both engines.

	30 OSD cluster (64K, 4K, mixed; latency, CPU, memory, capacity)	5 and 10 OSD clusters (4 MiB)
Load tool	fio 3.43, rados engine, stock librados, 3 client hosts x 16 jobs x QD16 = 768 I/Os in flight (QD1 cells: 48)	fio 3.43, rados engine, stock librados. Fill: 3 client hosts x 1 job x QD32. Load points: 1 stream (1 client host x 1 job x QD 1 / 4 / 16); 3 client hosts x 1 / 4 / 16 / 32 jobs x QD16 = 48 / 192 / 768 / 1,536 in flight
Pools	replicated size 3 / min_size 2, failure domain host, 1024 PGs; erasure coded 2+2 (isa), failure domain host, 512 PGs	replicated size 3 / min_size 2, failure domain host, 128 PGs (5 OSDs) / 256 PGs (10 OSDs)
Block sizes	64K and 4K (append, seq read, random write, random read), mixed 70 % read / 30 % write (bssplit 1k/10, 4k/30, 32k/30, 128k/20, 4m/10)	4M: sequential write (fill), random write, random read
Data per cell	identical for both engines: 4K 128 GiB to 640 GiB, 64K 1.25 TiB to 4.5 TiB, mixed 2 TiB; 4 MiB objects	fill writes the whole data set once (512 GiB on 5 OSDs, 1 TiB on 10 OSDs, 4 MiB objects, identical for both engines); random write and random read address that data set
Run time per cell	fixed data volume, not time-based; measured 55 to 237 s (BlueStore), 34 to 137 s (FlashStore); cap 600 s, none reached	load cells time based: 15 s ramp + 90 s measured (90 to 94 s BlueStore, 90 to 91 s FlashStore); fill: fixed volume; cap 600 s, none reached
Repeats	single run per cell and engine	single run per cell and engine
Cache and settle	BlueStore reference: no host cache drop between cells. FlashStore: before every cell all 30 OSDs restarted and the page cache of all OSD hosts dropped (cold start). Both: wait until the cluster is idle again (device writes, busy %, OSD CPU, in-flight ops at the idle baseline, 10 s windows)	before every load cell: page cache of all OSD hosts and the caches of every OSD dropped, then wait until the cluster is idle again (settle); not before the fill (empty pool)
Background work	scrub, deep scrub, balancer and PG autoscaler off; osd_op_queue wpq	same
BlueStore config	osd_memory_target 4 GiB (default, not overridden); no bluestore_* option set; discard off (bdev_enable_discard false)	osd_memory_target 4 GiB set on every OSD of both engines; no bluestore_* option set
Drive start state	both engines start on deallocated drives (blkdiscard)	both engines start on deallocated drives (blkdiscard)

The 5- and 10-OSD runs were done after the 30-OSD runs on the same hosts: the same drive model and CPU cores per OSD and the same client hosts, with smaller clusters because the 4 MiB cells of the 30-OSD cluster were bound by the client network.

Lower latency under load

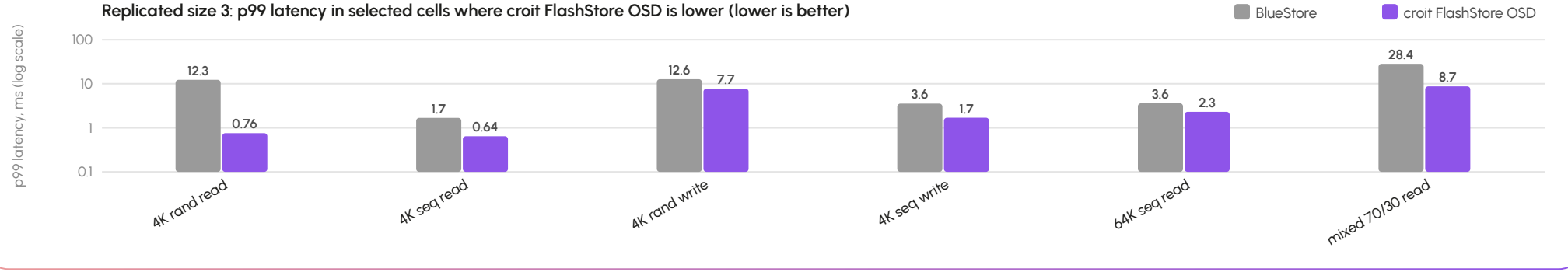
p99 / p50 = client completion latency under 48 clients x QD16 load, 30-OSD cluster.

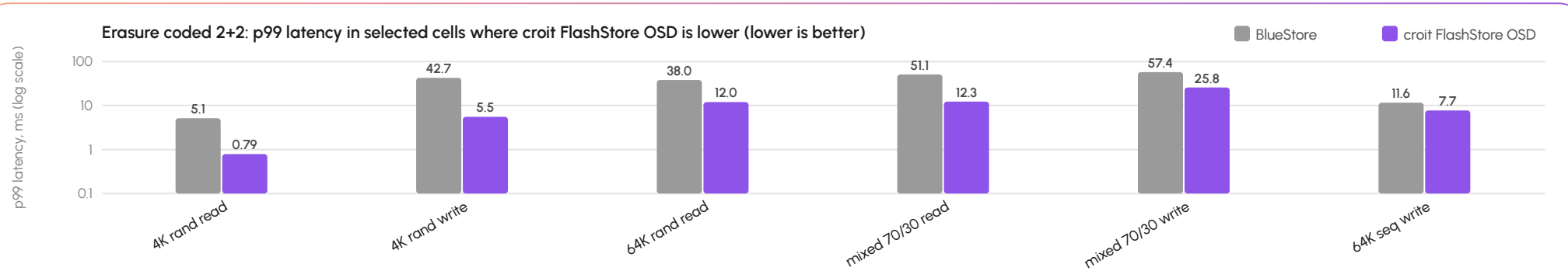
-52 %
p99 latency, replicated
up to -94 % (4K rand read)
Average over all 10 rep3 cells

-57 %
p99 latency, EC 2+2
up to -87 % (4K rand write)
Average over all 10 EC 2+2 cells

-23 %
Median latency (p50), replicated
up to -61 % (64K seq read)
Average over all 10 rep3 cells

-24 %
Median latency (p50), EC 2+2
up to -57 % (64K rand write)
Average over all 10 EC 2+2 cells



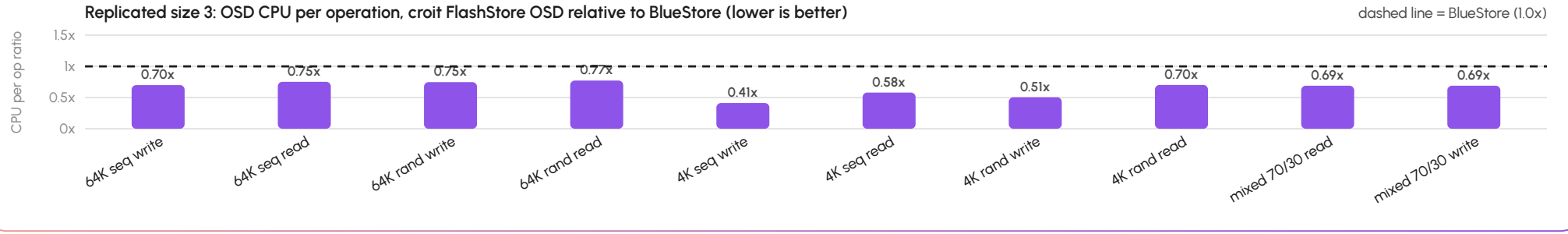


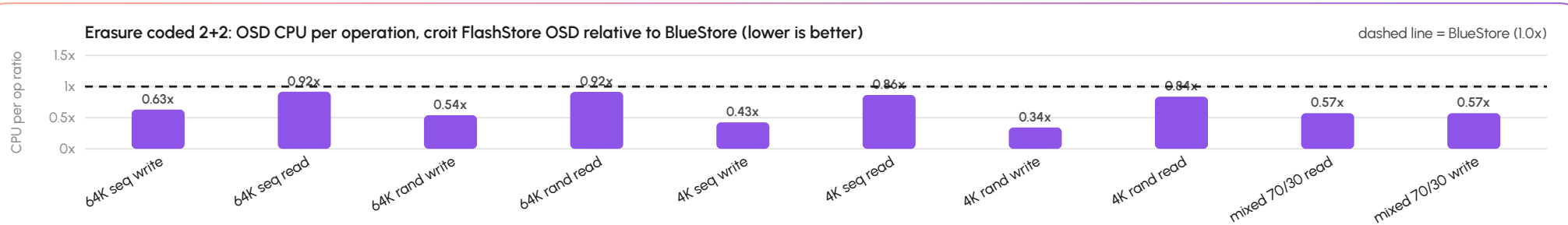
Charts show the selected cells named in their titles; the tiles above are averages over every cell of the family.

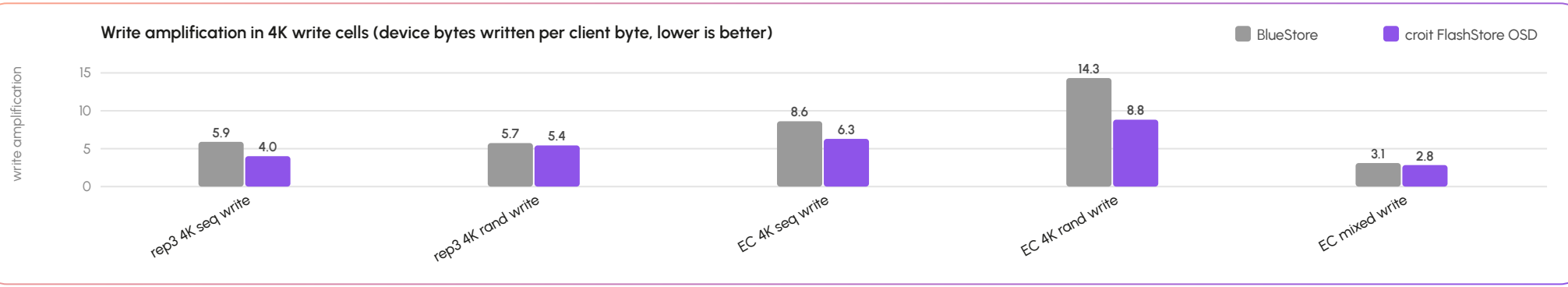
croit

36 % less OSD CPU per operation

OSD CPU microseconds per client operation and device write amplification.



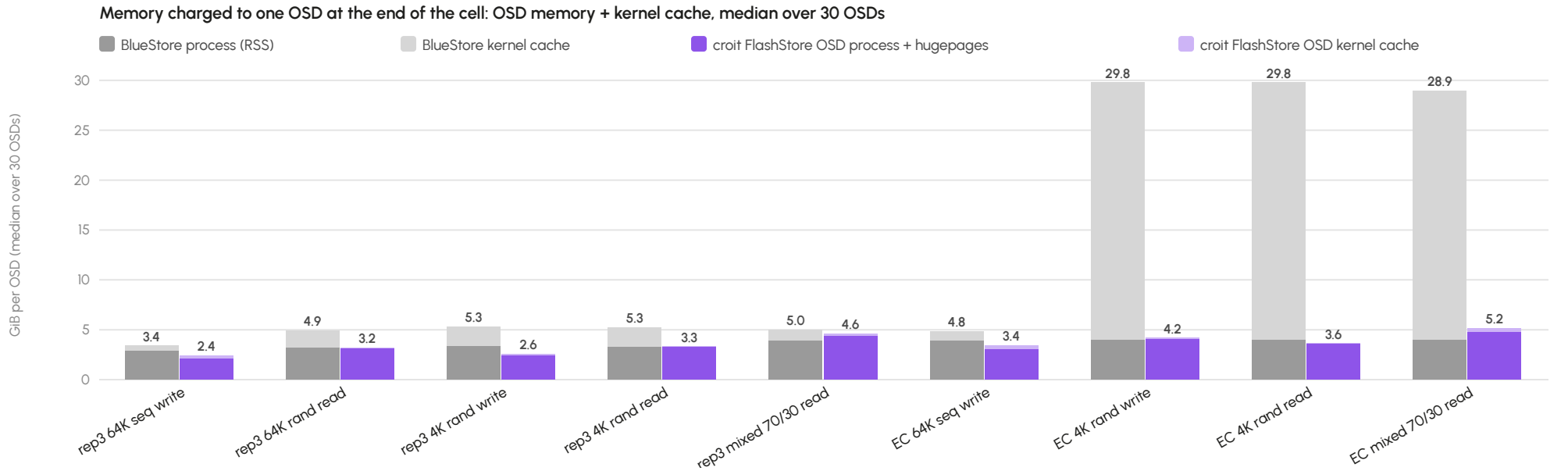




CPU per op: -36 % average (rep3), -37 % (EC 2+2); all 20 cells use less. Write amplification: 4K write cells shown.

Predictable memory: no growing block cache

Where the memory charged to an OSD goes (30-OSD cluster, median over 30 OSDs).



BlueStore cache = kernel block buffer cache charged to the OSD cgroup: BlueFS reads RocksDB through the page cache (bluefs_buffered_io, Ceph default, not changed). Its size follows the metadata of the data set of the cell: small random writes and EC 2+2 4K writes create far more RocksDB metadata than sequential writes, and the read cell keeps what the write cell built; the cache shrinks when the data set is removed before the next cell pair. The kernel can reclaim it. It reached up to 26 GiB per OSD (median over 30 OSDs). croit FlashStore OSD uses direct I/O and 2 MiB hugepages (included in its bar): its kernel cache charged to the OSD is at most 0.4 GiB and its memory per OSD stays between 2.2 and 4.8 GiB at cell end (median over OSDs) in every cell.

The FlashStore OSD memory (process + hugepages) is 13 % lower than the BlueStore process memory (median over 30 OSDs, geometric mean over 28 cells; BlueStore process 2.9-4.0 GiB). Both engines ran with osd_memory_target 4 GiB.

Capacity: on par

The space the same data occupies (30-OSD cluster).

18.93 TiB

Raw capacity used

BlueStore 19.09 TiB

Cluster raw used after the same test cycle, same data written

Raw capacity used after the test cycle (TiB, lower is better)



BlueStore croit FlashStore OSD

Cluster raw capacity used after the same test cycle: 18.93 TiB vs 19.09 TiB, on par (raw bytes used reported by the cluster at the end of the replicated and erasure coded 30-OSD cells, same data written by both engines). Single measurement.

croit

OMAP on one OSD: B+tree against BlueStore

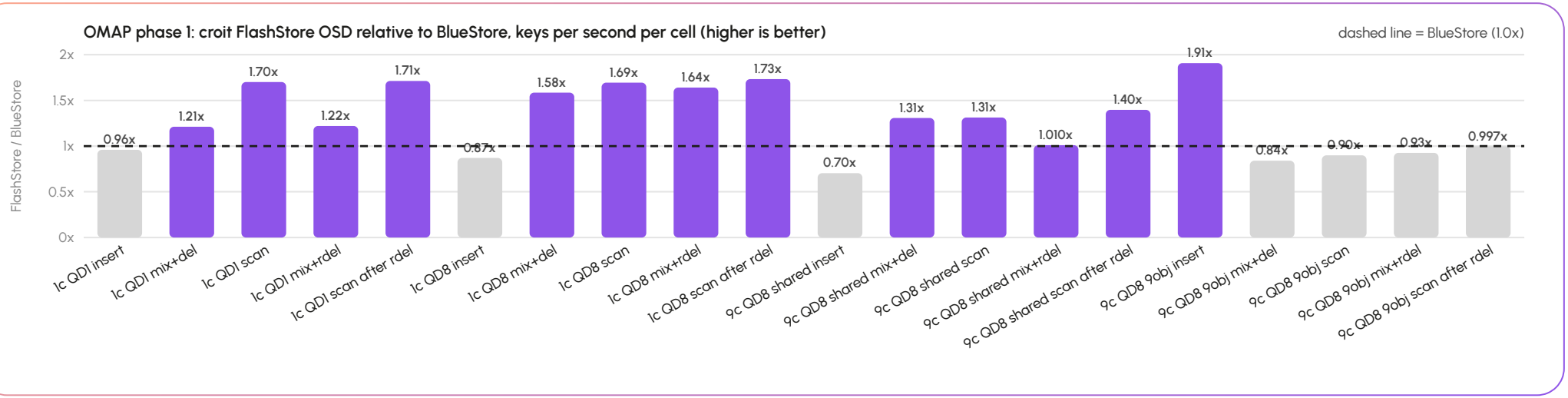
Phase 1: one OSD on one drive, replicated size 1, 1 or 9 clients. Keys per second, croit FlashStore OSD (B+tree OMAP engine) relative to BlueStore.

+23 %
OMAP throughput, average
geometric mean over 20 keys/s cells
insert 1.03x, mix 1.19x, scan 1.39x

13 of 20
Cells at or above BlueStore
7 cells below
ratio FlashStore / BlueStore, keys per second

1.91x
Best keys/s cell
9c QD8 9obj insert
BlueStore 1,856,830 keys/s, FlashStore 3,542,252 keys/s

0.70x
Lowest cell
9c QD8 shared insert
BlueStore 2,061,234 keys/s, FlashStore 1,450,713 keys/s



Tool and setup

- oprobe omapgate (same binary and commands for both engines); pool gate1: replicated size 1, 32 PGs, on one OSD (one NVMe drive)
- 1 client = one client host; 9 clients = 3 per client host on 3 client hosts
- Every op: 64 keys (one batch), 256 B values
- croit FlashStore OSD build 422be2684899 with the B+tree OMAP engine; BlueStore: Ceph 21.2.0 stock (RocksDB OMAP)

What the cells measure

- insert: keys into empty objects
- mix: equal parts insert, update, point read and delete (or range delete of 64 keys), a fixed 65,536 ops per client (per object in 72obj)
- scan: paged full key listing (1,024 keys per page) of the objects the mix left, repeated until at least 60 s

Reading the numbers

- keys/s = keys inserted, updated, read, deleted or listed per second, summed over all clients (delete objects: ops/s); ratio = FlashStore / BlueStore, above 1.0 is faster
- One run per cell and engine; FlashStore cells began with a restart of the OSDs and a page-cache drop, the BlueStore reference cells (2026-11-02) did not
- Grey bars and ratios below 1.0 are shown as measured

OMAP phase 1: measured values per cell

Phase 1: every cell of the chart, BlueStore and FlashStore values (single run each).

Clients, queue depth	Objects	Operation	BlueStore	FlashStore	FS / BS	unit	R
1 client, QD1	1 object	insert	353,415	339,229	0.96x	keys/s	
1 client, QD1	1 object	mixed with delete	206,467	250,205	1.21x	keys/s	
1 client, QD1	1 object	scan after delete mix	1,613,596	2,744,964	1.70x	keys/s	
1 client, QD1	1 object	mixed with range delete	230,566	281,269	1.22x	keys/s	
1 client, QD1	1 object	scan after range-delete mix	1,618,637	2,771,199	1.71x	keys/s	
1 client, QD8	1 object	insert	1,354,627	1,178,610	0.87x	keys/s	
1 client, QD8	1 object	mixed with delete	386,884	613,142	1.58x	keys/s	
1 client, QD8	1 object	scan after delete mix	1,613,993	2,735,108	1.69x	keys/s	
1 client, QD8	1 object	mixed with range delete	445,014	729,928	1.64x	keys/s	
1 client, QD8	1 object	scan after range-delete mix	1,610,354	2,790,642	1.73x	keys/s	
9 clients, QD8 each	shared objects	insert	2,061,234	1,450,713	0.70x	keys/s	
9 clients, QD8 each	shared objects	mixed with delete	643,567	841,660	1.31x	keys/s	
9 clients, QD8 each	shared objects	scan after delete mix	5,076,427	6,660,722	1.31x	keys/s	
9 clients, QD8 each	shared objects	mixed with range delete	935,021	944,096	1.01x	keys/s	
9 clients, QD8 each	shared objects	scan after range-delete mix	4,894,672	6,838,991	1.40x	keys/s	
9 clients, QD8 each	9 objects	insert	1,856,830	3,542,252	1.91x	keys/s	
9 clients, QD8 each	9 objects	mixed with delete	796,354	669,716	0.84x	keys/s	
9 clients, QD8 each	9 objects	scan after delete mix	8,705,294	7,829,675	0.90x	keys/s	
9 clients, QD8 each	9 objects	mixed with range delete	778,711	721,215	0.93x	keys/s	
9 clients, QD8 each	9 objects	scan after range-delete mix	8,840,146	8,812,103	1.00x	keys/s	

Keys per second summed over all clients (delete objects: operations per second); FS / BS above 1.0 (purple) = FlashStore faster, below 1.0 (grey) = BlueStore faster. Average (geometric mean over the 20 keys/s cells): 1.23x (+23 %). Chart bars above 2.0x are clipped, the value is printed.

croit

OMAP on 30 OSDs: B+tree against BlueStore

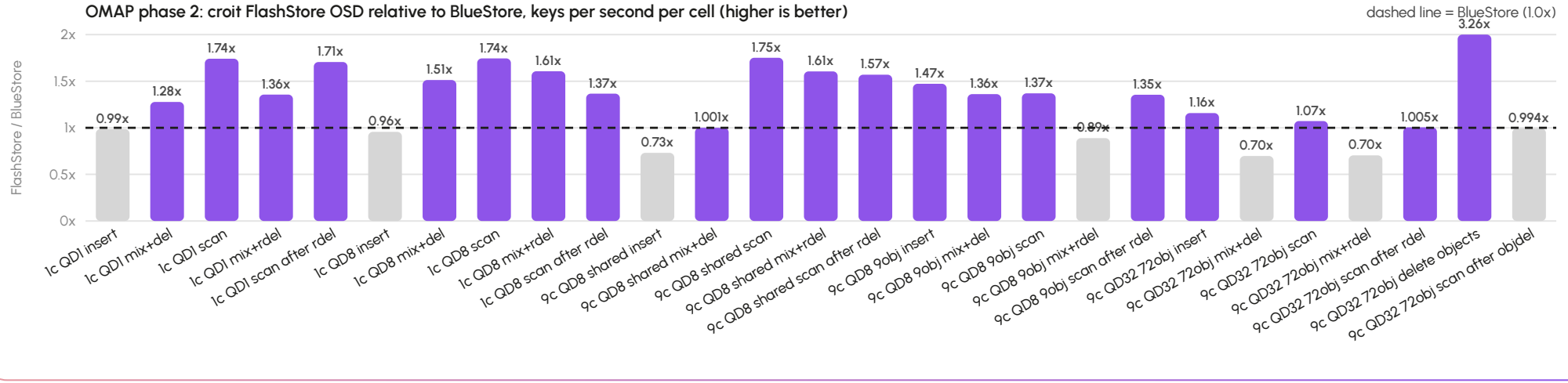
Phase 2: 30 OSDs, replicated size 3, 1 or 9 clients. Keys per second, croit FlashStore OSD (B+tree OMAP engine) relative to BlueStore.

+22 %
OMAP throughput, average
geometric mean over 26 keys/s cells
delete objects (ops/s) not averaged; insert 1.03x, mix 1.15x, scan 1.39x

19 of 26
Cells at or above BlueStore
7 cells below
ratio FlashStore / BlueStore, keys per second

1.75x
Best keys/s cell
9c QD8 shared scan after delete mix
BlueStore 4,675,280 keys/s, FlashStore 8,184,232 keys/s

0.70x
Lowest cell
9c QD32 72obj mixed with delete
BlueStore 5,634,853 keys/s, FlashStore 3,929,745 keys/s



Tool and setup

- oprobe omapgate (same binary and commands for both engines); pool gate2-rep3: replicated size 3, 1,024 PGs, failure domain host, 30 OSDs
- 1 client = one client host; 9 clients = 3 per client host on 3 client hosts; 72 objects = 8 per client
- Every op: 64 keys (one batch), 256 B values
- croit FlashStore OSD build 422be2684899 with the B+tree OMAP engine; BlueStore: Ceph 21.2.0 stock (RocksDB OMAP)

What the cells measure

- insert: keys into empty objects
- mix: equal parts insert, update, point read and delete (or range delete of 64 keys), a fixed 65,536 ops per client (per object in 72obj)
- scan: paged full key listing (1,024 keys per page) of the objects the mix left, repeated until at least 60 s
- delete objects / scan after object delete: 36 of 72 objects deleted with all keys, then the other 36 scanned (72obj)

Reading the numbers

- keys/s = keys inserted, updated, read, deleted or listed per second, summed over all clients (delete objects: ops/s); ratio = FlashStore / BlueStore, above 1.0 is faster
- One run per cell and engine; FlashStore cells began with a restart of the OSDs and a page-cache drop, the BlueStore reference cells (2026-11-02) did not
- Grey bars and ratios below 1.0 are shown as measured: 4 phase-2 cells were re-measured on 2026-11-08 (marked R in the table)

croit

OMAP phase 2: measured values per cell

Phase 2: every cell of the chart, BlueStore and FlashStore values (single run each).

Clients, queue depth	Objects	Operation	BlueStore	FlashStore	FS / BS	unit	R
1 client, QD1	1 object	insert	299,664	295,691	0.99x	keys/s	
1 client, QD1	1 object	mixed with delete	172,324	220,190	1.28x	keys/s	
1 client, QD1	1 object	scan after delete mix	1,595,527	2,777,686	1.74x	keys/s	
1 client, QD1	1 object	mixed with range delete	180,457	244,879	1.36x	keys/s	
1 client, QD1	1 object	scan after range-delete mix	1,608,208	2,742,533	1.71x	keys/s	
1 client, QD8	1 object	insert	846,901	809,762	0.96x	keys/s	
1 client, QD8	1 object	mixed with delete	300,759	454,881	1.51x	keys/s	
1 client, QD8	1 object	scan after delete mix	1,593,633	2,778,956	1.74x	keys/s	
1 client, QD8	1 object	mixed with range delete	329,843	529,700	1.61x	keys/s	
1 client, QD8	1 object	scan after range-delete mix	1,988,931	2,717,496	1.37x	keys/s	
9 clients, QD8 each	shared objects	insert	1,299,810	951,233	0.73x	keys/s	
9 clients, QD8 each	shared objects	mixed with delete	1,037,243	1,037,959	1.00x	keys/s	
9 clients, QD8 each	shared objects	scan after delete mix	4,675,280	8,184,232	1.75x	keys/s	
9 clients, QD8 each	shared objects	mixed with range delete	1,143,239	1,834,910	1.61x	keys/s	
9 clients, QD8 each	shared objects	scan after range-delete mix	5,134,411	8,058,616	1.57x	keys/s	
9 clients, QD8 each	9 objects	insert	3,755,069	5,529,521	1.47x	keys/s	
9 clients, QD8 each	9 objects	mixed with delete	1,662,951	2,264,017	1.36x	keys/s	
9 clients, QD8 each	9 objects	scan after delete mix	10,336,074	14,165,293	1.37x	keys/s	
9 clients, QD8 each	9 objects	mixed with range delete	2,097,046	1,862,744	0.89x	keys/s	
9 clients, QD8 each	9 objects	scan after range-delete mix	10,704,417	14,492,416	1.35x	keys/s	
9 clients, QD8 / QD32	72 objects	insert	6,959,599	8,061,738	1.16x	keys/s	
9 clients, QD8 / QD32	72 objects	mixed with delete	5,634,853	3,929,745	0.70x	keys/s	
9 clients, QD8 / QD32	72 objects	scan after delete mix	14,755,087	15,830,281	1.07x	keys/s	
9 clients, QD8 / QD32	72 objects	mixed with range delete	6,016,751	4,237,650	0.70x	keys/s	R
9 clients, QD8 / QD32	72 objects	scan after range-delete mix	15,062,389	15,132,004	1.00x	keys/s	R
9 clients, QD8 / QD32	72 objects	delete 36 objects	92	299	3.26x	ops/s	R
9 clients, QD8 / QD32	72 objects	scan after object delete	14,483,993	14,392,411	0.99x	keys/s	R

Keys per second summed over all clients (delete objects: operations per second); FS / BS above 1.0 (purple) = FlashStore faster, below 1.0 (grey) = BlueStore faster. Average (geometric mean over the 26 keys/s cells, delete objects excluded): 1.22x (+22 %). Chart bars above 2.0x are clipped, the value is printed. R = re-measured on 2026-11-08 with the benchrun harness (missing in the 2026-11-07 run): keys of all clients / (last end - first start); delete objects counts delete and verify-absent operations like the BlueStore reference.

croit

Rust client libraries against stock librados and librbd

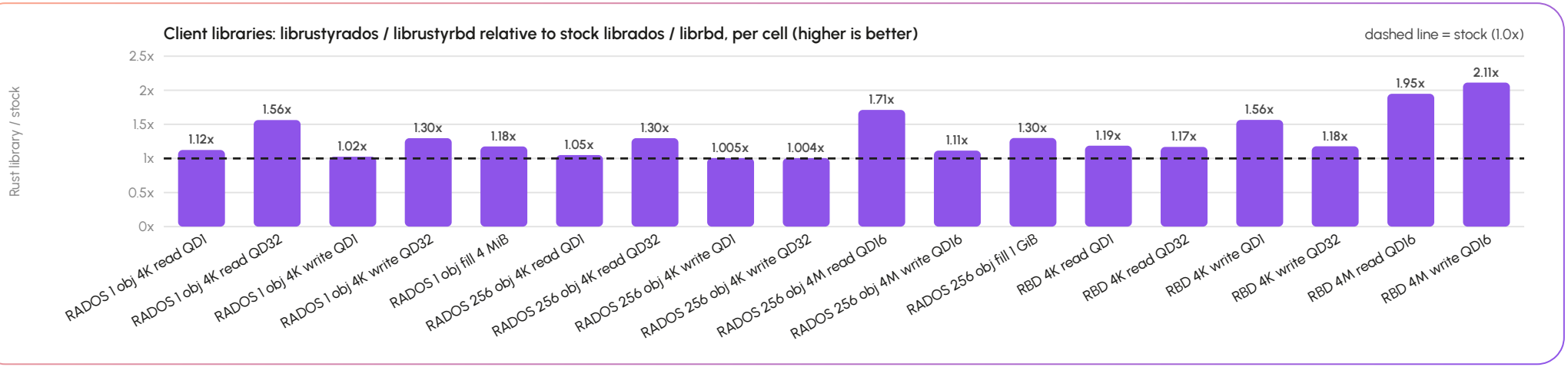
librustyrados and librustyrbd are drop-in replacements for librados and librbd, written in Rust. Same cluster and fio binary, one client host, 18 cells per library.

2.1x
RBD 4M writes, one client
librustyrbd 17.9 vs 8.5 GB/s
QD16, replicated size 3, cache ack vs cache ack (durable at flush)

1.9x
RBD 4M reads, one client
librustyrbd 20.2 vs 10.4 GB/s
one image, QD16

1.7x
RADOS 4M reads, one client
librustyrados 18.3 vs 10.7 GB/s
256 objects of 4 MiB, QD16

1.6x
RADOS 4K reads of one object
librustyrados 136k vs 87k IOPS
QD32, one object (one PG, one OSD session)



Setup

- Client host: one of the 8 hosts (AMD EPYC 9455P, 48 threads, 2 x 200 Gbit/s); OSD cluster: 5 hosts, 30 NVMe OSDs, 30 croit FlashStore OSDs (6 per host on hosts 1-5, kernel NVMe tier, build 422be2684899), Ceph 21.2.0, kernel TCP
- Libraries: stock Ceph 21.2.0 librados/librbd against librustyrados/librustyrbd (ceph-rust build 8ace974c4e)
- fio 3.43, rados and rbd engines, the same fio binary for both; only the loaded library differs

Cells

- Replicated size 3, 512 PGs, one pool per library; one client host, one fio job, 60 s per cell
- 18 cells per library, each run once per library; RBD: one 1 GiB image per cell, RADOS: 1 or 256 objects of 4 MiB
- No OSD restart or cache drop before the cells; fill cells measure one 4 MiB write (latency) and a 1 GiB write (wall time)

Reading the numbers

- Ratio = librustyrados or librustyrbd IOPS / stock IOPS (fill cells: time); above 1.0 = Rust library faster
- Faster than or equal to stock in all 18 measured cells; the 256-object 4K random write cells (QD1, QD32) are parity (1.005x, 1.004x), bound by the OSD cluster (both libraries)
- RBD write cells compare cache acknowledgement with cache acknowledgement: both libraries use the librbd default write-around cache, and data is durable at flush.
- librustyrados and librustyrbd are available as a preview for testing and evaluation in the croit frontier release. They are not yet recommended for production data.

Already running in our labs: Ceph, reshaped for AI

The croit FlashStore OSD is the first milestone. Next, the same fast path runs all the way to the client, to virtual machines and to GPUs: from the drive to the GPU.

1

librustyrados

Preview in the frontier release

librustyrados and librustyrbd are available as a preview for testing and evaluation in the croit frontier release. They are not yet recommended for production data. Measured against stock librados: page 17.

2

librustyrbd

Block storage on the new client

RBD images for KVM/QEMU on top of librustyrados; existing images and tools stay compatible. Same preview status; measured against stock librbd: page 17.

3

RDMA

Between OSDs and clients

RoCEv2 between OSDs for replication and recovery, and between librusty* clients and OSDs: data moves NIC to NIC without the TCP stack. Standard Ethernet, opt-in.

4

KVM over RDMA

Straight into the OSD

VM disk I/O travels over RDMA from the hypervisor into the OSD, with no TCP in the data path: lower latency for demanding virtual machines.

5

GPUDirect for AI

NIXL and KVBM

Through librustyrados, NVIDIA NIXL and the Dynamo KV Block Manager (KVBM) move KV cache and model data directly between GPU memory and the OSD, without staging through host memory.

librustyrados, librustyrbd and RDMA are in public preview since 9 November 2026; the KVBM/NIXL plugin follows as a public preview in Q1 2027. All dates are indicative and subject to change based on croit customer demand.

croit

Measured values per cell

Replicated size 3, 30-OSD cluster, 48 clients x QD16: all 10 cells (4 MiB: pages 5 and 6)

Cell	IOPS			p50 ms			p99 ms			CPU µs/op			OSD memory GiB (median)			Write amp.	
	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS
64K seq write	375,783	400,838	**+7 %**	1.63	1.60	**-2 %**	7.24	5.67	**-22 %**	256	179	**-30 %**	2.9	2.2	**-25 %**	3.2	3.1
64K seq read	522,622	990,806	**+90 %**	1.61	0.63	**-61 %**	3.59	2.31	**-36 %**	32	24	**-25 %**	2.9	3.1	+6 %	-	-
64K rand write	329,119	438,560	**+33 %**	1.25	1.14	**-9 %**	16.32	11.08	**-32 %**	370	277	**-25 %**	3.1	2.5	**-22 %**	3.2	3.2
64K rand read	1,060,164	1,045,845	-1 %	0.49	0.45	**-7 %**	2.61	3.16	+21 %	46	36	**-23 %**	3.2	3.2	**-1 %**	-	-
4K seq write	612,167	909,481	**+49 %**	1.09	0.79	**-27 %**	3.56	1.70	**-52 %**	157	65	**-59 %**	3.3	2.3	**-30 %**	5.9	4.0
4K seq read	989,827	2,209,461	**+123 %**	0.72	0.32	**-56 %**	1.68	0.64	**-62 %**	18	11	**-42 %**	3.3	3.3	**-2 %**	-	-
4K rand write	470,596	979,347	**+108 %**	0.85	0.49	**-42 %**	12.65	7.70	**-39 %**	246	125	**-49 %**	3.4	2.5	**-27 %**	5.7	5.4
4K rand read	1,450,914	2,380,557	**+64 %**	0.20	0.28	+37 %	12.26	0.76	**-94 %**	35	24	**-30 %**	3.3	3.3	**-2 %**	-	-
mixed 70/30 read	334,981	471,237	**+41 %**	0.34	0.32	**-7 %**	28.44	8.72	**-69 %**	210	145	**-31 %**	3.9	4.4	+11 %	-	-
mixed 70/30 write	143,576	201,957	**+41 %**	1.09	0.98	**-10 %**	34.87	27.13	**-22 %**	210	145	**-31 %**	3.9	4.4	+11 %	3.9	3.6

These cells feed: throughput, p99, p50 and CPU averages (replicated). BS = BlueStore, FS = croit FlashStore OSD, Δ = change of FS vs BS (purple = better, grey = worse). OSD memory = process RSS (FlashStore: plus its hugepage share) per OSD at the end of the cell, median over 30 OSDs. Write amplification = device bytes written per client byte, write cells only. Mixed 70/30 cells are listed per direction.

croit

Measured values per cell

Erasure coded 2+2, 30-OSD cluster, 48 clients x QD16: all 10 cells

Cell	IOPS			p50 ms			p99 ms			CPU µs/op			OSD memory GiB (median)			Write amp.	
	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS
64K seq write	259,503	319,737	**+23 %**	2.31	1.99	**-14 %**	11.60	7.70	**-34 %**	382	241	**-37 %**	3.9	3.1	**-21 %**	2.3	2.2
64K seq read	371,786	552,945	**+49 %**	1.83	1.12	**-39 %**	6.32	5.21	**-18 %**	104	95	**-8 %**	3.8	4.0	+5 %	-	-
64K rand write	147,056	246,991	**+68 %**	3.23	1.38	**-57 %**	40.63	33.16	**-18 %**	888	480	**-46 %**	3.9	3.8	**-1 %**	2.5	2.5
64K rand read	594,239	711,650	**+20 %**	0.49	0.54	+10 %	38.01	11.99	**-68 %**	136	125	**-8 %**	4.0	3.8	**-6 %**	-	-
4K seq write	324,304	498,078	**+54 %**	2.01	1.38	**-31 %**	6.78	3.85	**-43 %**	319	136	**-57 %**	4.0	2.8	**-30 %**	8.6	6.3
4K seq read	1,043,835	1,150,923	**+10 %**	0.66	0.54	**-17 %**	1.71	2.34	+37 %	42	37	**-14 %**	4.0	3.8	**-5 %**	-	-
4K rand write	141,414	405,178	**+187 %**	3.29	1.66	**-50 %**	42.73	5.54	**-87 %**	957	327	**-66 %**	4.0	4.1	+1 %	14.3	8.8
4K rand read	1,378,662	1,901,424	**+38 %**	0.36	0.37	+2 %	5.14	0.79	**-85 %**	65	54	**-16 %**	4.0	3.6	**-10 %**	-	-
mixed 70/30 read	164,094	285,006	**+74 %**	0.83	0.86	+4 %	51.12	12.26	**-76 %**	503	288	**-43 %**	4.0	4.8	+21 %	-	-
mixed 70/30 write	70,275	122,161	**+74 %**	2.28	1.86	**-18 %**	57.41	25.82	**-55 %**	503	288	**-43 %**	4.0	4.8	+21 %	3.1	2.8

These cells feed: throughput, p99, p50 and CPU averages (EC 2+2). BS = BlueStore, FS = croit FlashStore OSD, Δ = change of FS vs BS (purple = better, grey = worse). OSD memory = process RSS (FlashStore: plus its hugepage share) per OSD at the end of the cell, median over 30 OSDs. Write amplification = device bytes written per client byte, write cells only. Mixed 70/30 cells are listed per direction.

croit

Measured values per cell

Replicated size 3, 30-OSD cluster, queue depth 1: all 8 cells

Cell	IOPS			p50 ms			p99 ms			CPU µs/op			OSD memory GiB (median)			Write amp.	
	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS	Δ	BS	FS
QD1 64K seq write	130,265	147,013	**+13 %**	0.33	0.29	**-10 %**	0.82	0.54	**-34 %**	333	209	**-37 %**	3.9	2.5	**-37 %**	3.3	3.2
QD1 64K seq read	240,595	245,888	**+2 %**	0.18	0.18	**-2 %**	0.30	0.31	+4 %	35	33	**-6 %**	4.0	3.4	**-14 %**	-	-
QD1 64K rand write	118,678	131,953	**+11 %**	0.35	0.32	**-8 %**	0.89	0.70	**-21 %**	399	250	**-37 %**	3.9	2.8	**-30 %**	3.3	3.3
QD1 64K rand read	226,756	239,822	**+6 %**	0.20	0.19	**-5 %**	0.31	0.31	**-3 %**	47	36	**-22 %**	4.0	3.3	**-19 %**	-	-
QD1 4K seq write	169,167	200,011	**+18 %**	0.26	0.23	**-10 %**	0.62	0.43	**-30 %**	247	135	**-46 %**	4.0	2.3	**-41 %**	7.9	6.2
QD1 4K seq read	495,156	499,401	**+1 %**	0.07	0.07	**-1 %**	0.18	0.18	**-1 %**	23	22	**-6 %**	3.9	3.3	**-15 %**	-	-
QD1 4K rand write	159,316	188,199	**+18 %**	0.26	0.22	**-14 %**	0.75	0.55	**-26 %**	289	160	**-44 %**	3.8	2.5	**-34 %**	7.7	6.6
QD1 4K rand read	309,851	334,781	**+8 %**	0.14	0.13	**-9 %**	0.25	0.24	**-4 %**	40	27	**-32 %**	3.8	3.3	**-15 %**	-	-

These cells feed: QD1 IOPS average. BS = BlueStore, FS = croit FlashStore OSD, Δ = change of FS vs BS (purple = better, grey = worse). OSD memory = process RSS (FlashStore: plus its hugepage share) per OSD at the end of the cell, median over 30 OSDs. Write amplification = device bytes written per client byte, write cells only.

croit

TEST METHOD

- 8 hosts (AMD EPYC 9455P): 5 OSD hosts x 6 OSDs = 30 NVMe OSDs (Samsung 15.36 TB), 3 client hosts; 4 MiB cells on 5 OSDs (one per OSD host) and 10 OSDs (two per OSD host)
- Network: bond0 2 x 200 Gbit/s (LACP), MTU 1500, client and cluster network are the same; kernel TCP, no RDMA
- Load: fio rados engine with stock librados; replicated size 3 and erasure coded 2+2 pools; 3 client hosts x 16 jobs x QD16
- BlueStore: Ceph 21.2.0 stock; croit FlashStore OSD: build 422be2684899 on all clusters, Ceph 21.2.0 compatible; osd_memory_target 4 GiB for both
- Same cells and data volumes for both arms; cluster settled to idle before every cell; single run per cell
- Averages are geometric means of the per-cell ratio croit FlashStore OSD / BlueStore over the stated cell family; 'up to' is the best single cell
- OSD memory = process RSS (+ hugepage share for FlashStore) at the end of the cell, median over 30 OSDs; kernel cache = cgroup memory minus RSS; BlueStore reference measured 2026-11-02 (30 OSDs) and 2026-11-03 (4 MiB), croit FlashStore OSD measured 2026-11-07, same cluster, same test specification; full conditions on pages 7 and 8



Keep it yours.

Storage, virtualization and AI, all in house.

Version 09 Nov 2026 17:01