PRESS Q&A

croit FlashStore OSD: questions and answers

For editors. Background and answers to the questions we expect.
croit FlashStore OSD is available as a preview for testing and evaluation. It is not yet recommended for production data.

> Figures refer to the benchmark report croit FlashStore OSD vs BlueStore (version 09 Nov 2026).

TECHNOLOGY

What is croit FlashStore OSD?
A Ceph object storage daemon (OSD) written from scratch in Rust for NVMe flash. It replaces BlueStore on a drive and talks to monitors, managers, clients and classic OSDs with the standard Ceph protocol, so it runs inside an existing cluster.

Why rewrite the OSD at all?
On flash, Ceph is limited by the CPU time the OSD spends per operation, not by the drive. Shared queues with locks, RocksDB and BlueFS under every object and the kernel paths all add up. FlashStore OSD removes these steps: one pinned thread per core, its own storage engine and a short io_uring path to the drive.

Why Rust?
Memory safety without a garbage collector, predictable latency and good support for io_uring and per-core designs. For a daemon that holds customer data, ruling out whole classes of memory bugs at compile time matters.

What does clean room mean here?
Written from scratch, without reusing Ceph's C++ code; it implements the Ceph wire protocol and behaviour from our own specifications.

Which data services does it support?
Replicated and erasure-coded pools. RBD and CephFS run on top, and the object classes RGW needs are already implemented; they use the OSD through the normal protocol.

Which Ceph versions work?
Ceph v21.2.0 (Umbrella), v20.2.5 and v20.2.4 (Tentacle); benchmarked on 21.2.0.

NUMBERS AND METHOD

What are the headline numbers?
Against Ceph 21.2.0 BlueStore on 30 NVMe OSDs: +50 % throughput on replicated pools and +54 % on erasure-coded 2+2 pools on average, -52 % p99 latency (replicated), -36 % OSD CPU per operation, and 2.2-4.8 GiB median memory per OSD. Best single cells: +123 % (4K seq read, replicated) and +187 % (4K rand write, EC 2+2).

How was it measured?
5 OSD hosts with 6 NVMe OSDs each (AMD EPYC 9455P, 6 dedicated cores per OSD), 3 client hosts, 2 x 200 Gbit/s, MTU 1500, kernel TCP, no RDMA. Load: fio 3.43 rados engine with stock librados, 48 clients x QD16, identical data volumes for both engines. Full conditions are on pages 7 and 8 of the benchmark report.

How does OMAP perform (RGW index, CephFS metadata)?
B+tree OMAP engine against BlueStore (RocksDB), 64-key batches. On 30 OSDs (replicated size 3) FlashStore OSD is +22 % on average over 26 keys/s cells, at or above BlueStore in 19; on one OSD +23 % over 20 cells, 13 at or above. All cells, losses included: report pages 13 to 16.

Why RADOS and not RBD or a file system?
The benchmark measures the OSD, which is the component that changed. RADOS load removes client-side effects. Whether RBD, RGW or CephFS workloads get faster depends on whether the OSD is their bottleneck. The OSD figures are therefore RADOS figures. The Rust client libraries are measured separately, see the next answer.

Why are 4 MiB objects shown on 5 and 10 OSDs?
On 30 OSDs the 4 MiB cells were limited by the client network (69 to 73 GB/s on both engines). That says nothing about the OSD, so we measured 4 MiB on smaller clusters where the OSDs are the limit: writes +14 % (5 OSDs) and +28 % (10 OSDs), reads +1 % and +4 %.

How often was each cell run?
Each cell runs once per engine. The raw fio data of every run is available on request as a redacted data kit.

Did the figures change, and will they change?
They can. FlashStore OSD is a young product and we keep optimising it, so newer builds will measure differently, and we expect them to improve. Every report carries its version date and build IDs; the figures here refer to the report of 09 Nov 2026. We will publish updated results with their date and conditions.

Can I check the results myself?
Yes. On request we provide the raw fio output of all cells, the job files and the cluster configuration as a redacted data kit. Hands-on testing is supervised: you tell our engineers what to test, and they run it live in front of you.

WHERE IT IS WEAKER

Where is FlashStore OSD slower than BlueStore?
IOPS is at or above BlueStore in 27 of 28 cells (replicated 64K rand read: 0.99x), and p99 latency is lower in 25 of 28 cells. The cells where it is not lower: replicated 64K rand read, EC 2+2 4K seq read, QD1 replicated 64K seq read. All values are in the report.

Are there other known weak spots?
Yes, in OMAP (key/value) workloads: on 30 OSDs 4 of 27 cells are more than 10 % below BlueStore: 9c QD8 shared insert (0.73x); 9c QD8 9obj mixed with range delete (0.89x); 9c QD32 72obj mixed with delete (0.70x); 9c QD32 72obj mixed with range delete (0.70x); on one OSD 4 of 20: 1c QD8 insert (0.87x); 9c QD8 shared insert (0.70x); 9c QD8 9obj mixed with delete (0.84x); 9c QD8 9obj scan after delete mix (0.90x). Both are being worked on.

Does it use more memory?
No: in total it uses far less, because it does not fill the kernel page cache: 2.2-4.8 GiB median per OSD at cell end, against up to 26 GiB of kernel cache per OSD for BlueStore in EC 4K cells. The OSD memory itself (process plus hugepages) is about 13 % lower than the BlueStore process memory.

What about HDDs?
FlashStore OSD is built for NVMe flash only. There is no HDD path and none is planned. On spinning disks the drive is the limit, which is where BlueStore belongs.

COMPARISON

Why no comparison with Crimson/SeaStore?
Crimson with SeaStore is the upstream next-generation OSD and 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.

How does this relate to IBM Storage Ceph or other Ceph vendors?
FlashStore OSD is wire compatible with Ceph, so any Ceph cluster on NVMe can run it next to classic OSDs. croit customers get it integrated in the croit platform.

Will it go upstream?
Yes. We want FlashStore OSD to become part of the Ceph ecosystem and will contribute upstream. Our goal is that the whole Ceph community can move to it.

OPERATIONS

How do I migrate?
There is no in-place conversion, because the on-disk format differs. Drives are redeployed as FlashStore OSDs and backfilled from the other copies. Depending on the cluster setup and its failure domains, this runs one drive or one whole server at a time. Classic and FlashStore OSDs serve the same pool throughout, and you can stop after any step. Erasure-coded pools: same procedure.

Can I go back to BlueStore?
Yes, the same way: redeploy the drive as a BlueStore OSD and let the cluster backfill it. It takes time on large clusters, but it is the standard Ceph procedure.

Is it production ready?

croit FlashStore OSD is available as a preview for testing and evaluation. It is not yet recommended for production data. croit customers who test it get help from croit. General availability is planned for the first quarter of 2027. We will release it when our Ceph engineers are confident it is ready for production data, not by a fixed date.

What hardware do I need?

NVMe drives with power-loss protection, 4 KiB sector format (512 B namespaces must be reformatted) and Linux 5.14 or newer (RHEL / Alma 9) with io_uring enabled.

How do I enable it?

In croit settings (Settings, Add, value on): `croit enable frontier image` and `croit enable flashstore`. Then per drive: Disks, Set to OSD, Implementation: FlashStore.

CLIENT LIBRARIES

How fast are the Rust client libraries?

Faster than or equal to stock Ceph 21.2.0 librados and librbd in all 18 measured cells: RBD 4 MiB writes 2.1x (17.9 vs 8.5 GB/s) and reads 1.9x (20.2 vs 10.4 GB/s), RADOS 4 MiB reads 1.7x, RADOS 4K random reads of one object at QD32 1.6x, RBD 4K random writes at QD1 +56 %. RADOS 4K random writes over 256 objects are at parity: bound by the OSD cluster (both libraries). Method: one client host, one fio 3.43 job, 60 s per cell, each cell once per library, replicated size 3, the same 30-OSD croit FlashStore cluster (5 OSD hosts). RBD write cells compare cache acknowledgement with cache acknowledgement: both libraries use the librbd default write-around cache, and data is durable at flush. All cells: fact sheet page 4, report page 17. 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.

LICENCE AND BUSINESS

Under which licence is it released?

Open source under LGPL-3.0. Code and deployment guide: github.com/croit/ceph-rust; benchmark scripts and raw data in the data kit (on request).

What does it cost?

Nothing extra. FlashStore OSD is included in the croit frontier release of the croit platform. Anyone else can deploy it from the repository.

How does croit make money with it?

croit sells its Ceph platform with support, training and consulting. A faster OSD makes that platform more attractive, it is not sold separately.

How was it developed?

With a heavily AI-driven, test-driven workflow. A small senior Ceph team at croit, Martin Verges, Igor Fedotov and Md Mahamudur Rahaman Sajib, sets the design and oversees the project. AI agents write, review and test the changes, and every change has to pass an extensive test suite and benchmarks against BlueStore.

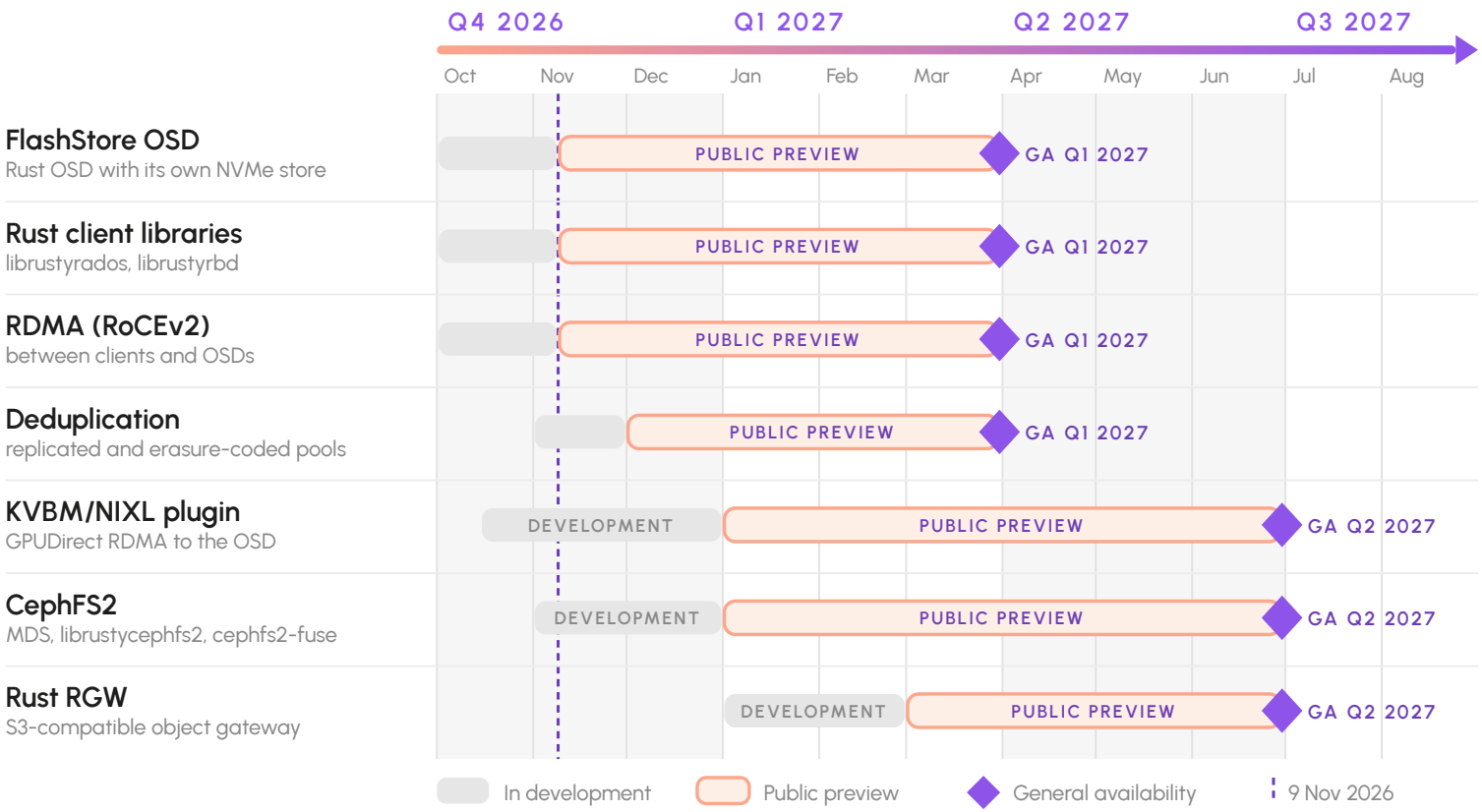
What comes next?

FlashStore OSD and the Rust client libraries librustyrados and librustyrbd with RDMA support are in public preview since 9 November 2026; general availability is planned for Q1 2027. Deduplication for replicated and erasure-coded pools follows as a public preview in December 2026, general availability in Q1 2027. In Q1 2027 come the public previews of the KVBM/NIXL plugin for GPUDirect RDMA to the OSD, of CephFS2, a transaction-based file system compatible with libcephfs (MDS, librustycephfs2 and cephfs2-fuse), and of a Rust RGW; their general availability is planned for Q2 2027. All dates are indicative and subject to change based on croit customer demand. The timeline is on page 4.

croit

Roadmap

Public previews from November 2026, general availability in Q1 and Q2 2027. What croit plans next around FlashStore OSD: native Rust clients, RDMA and GPUDirect, a new file system, deduplication and a Rust RGW.



All dates are indicative and subject to change based on croit customer demand.