



One Drive Sets the Tail

How a degraded SSD stalled a replicated Ceph tier, misled our incident system, and outlived its own repair

Paper: PT-R-2026-035 v1.0

Author: PureTensor Ltd

Date: 8 October 2026

Status: Published

Abstract

Our flash storage tier is four consumer SATA SSDs, one in each of four storage hosts, running as a single Ceph 19.2.6 device class. For several days the whole tier showed recurring multi-second commit stalls, and every OSD in the class looked equally slow in its slow-op history. On 30 September 2026 at 15:35:12 UTC the stalls surfaced as a bus-wide NATS JetStream timeout. Our autonomous incident system then filed a severity-3 incident against a compute node that was healthy throughout. We wanted to know whether the cause was the drive class, something shared, or one device.

The usual checks looked acceptable. These covered isolated benchmarks, SMART error counters, placement balance, network retransmits and one-hour device latency. A six-minute capture of all four drives at one-second resolution, joined second by second to Ceph's op history, gave a different answer. One drive, which we call B, reached a write-await p99 of 353.67 ms and a maximum of 725.75 ms. The p99 on its siblings was 7.23 to 9.44 ms. B stalled for several seconds at a time at 100 % utilisation, and slow ops on all four OSDs began in the same seconds. A replicated or erasure-coded write is acknowledged only when every member has committed, so the healthy drives' slow replica writes, near three seconds each, were waiting on B. A ten-minute re-measurement on 3 October reproduced the shape: B had 15 seconds over 100 ms and each sibling had 0. B's cache flushes reached 1,072 ms against about 17 ms on the others.

The long-run record places the change on 25 September, the day we moved metadata and erasure-coded pools onto this tier. B's calendar-day mean commit latency stepped from 9.3 to 11.5 ms to between 62.6 and 89.7 ms and stayed there. The tier mean was about 32 ms, which reads as a modestly slow tier rather than one broken drive. Nothing alerted for five days. SMART data shows that B barely uses its SLC cache: 2,248 GB written to SLC over its life, against 20,120 to 37,663 GB on its siblings. In a clean interval B wrote about 3.5 times the TLC NAND of each sibling for the same host data.

We rebuilt B in place, including a whole-device discard. During backfill its p99 was 2.73 ms. The fix held about ten hours after the OSD was recreated, and B's hourly mean commit latency has since ranged from 38.8 to 325.6 ms. The drive has not been replaced; a replacement is pending an operator decision. A commit-latency alert we added on 30 September is firing, accurately. The claim we can support is that one drive sets the latency tail of the tier and of the services above it. Why that drive degraded is not established.



1. The Tier and the Question

The cluster has four storage hosts running Proxmox VE on 25 GbE. Each host carries three enterprise HDD OSDs, which form the bulk tier. Each also carries exactly one SATA SSD OSD (an OSD is Ceph's object storage daemon, one per device), so the SSD device class is four OSDs, one per host. We call them A, B, C and D.

All four SSDs are WD Red SA500 2.5" 2 TB drives on the same firmware (540500WD). They run at SATA 6.0 Gb/s with write cache on, queue depth 32 and TRIM supported. They are consumer drives without power-loss protection (PLP, the onboard energy reserve that lets a drive acknowledge writes held in volatile cache). Without PLP, every synchronous write and every cache flush must reach NAND before the drive acknowledges it. BlueStore, Ceph's on-device storage backend, is colocated on each SSD. Data, the RocksDB metadata store and the write-ahead log (WAL) have shared the one device since the re-architecture of 25 September. Power-on hours are A 4,354, B 4,587, C 3,872 and D 2,432.

The SSD class serves two kinds of pool, both with host as the failure domain.

The first kind is erasure-coded with $k=3$ data chunks, $m=1$ parity chunk and `min_size 3`. Erasure coding stores each object as data chunks plus parity chunks, spread across hosts. One such pool holds CephFS data and RBD volumes for monitoring and other small services (101 GiB); the other is nearly empty. With $k+m=4$ on four hosts, every write to these pools touches every SSD, B included. One OSD down leaves zero redundancy margin.

The second kind is replicated, with `size 3` and `min_size 2`. These pools hold CephFS metadata (including the metadata server journal), an RBD pool and two near-empty pools. The RBD pool holds the NATS JetStream volumes and, since 27 September, the git forge's SQLite database. About three quarters of replicated placement groups (PGs, the units into which Ceph divides a pool and maps onto OSDs) include B. Each SSD OSD carries 481 to 488 PGs.

Load per SSD is about 85 to 110 writes/s and about 35 cache flushes/s. B's write rate tracks its siblings'. In one set of hourly node-exporter means, for example, B ran at 90.2 writes/s and A at 94.3. B was not receiving more work than the others.

2. The Symptom Above Ceph

NATS JetStream runs three replicas, with file storage on the replicated SSD RBD pool. At 15:35:12 on 30 September every probe of our autonomous incident system, on four nodes, failed to publish with "nats: timeout". Seven probe pods ended in the Failed state between 15:45 and 16:00, and they had cleared by 16:15.

At 15:35:13 NATS logged six "Internal subscription ... took too long" warnings for stream-info requests, each lasting 5.42 to 6.29 s, and two consumer-info waits of 2.21 s. Leaders re-elected at 15:35:15 to 15:35:16. At 15:40:29 it logged "Metalayer async snapshot took 4.610s ... compacted: 1.90 KB": 4.6 s to write 1.9 KB. This was not new. In the NATS pods' first 46 hours, the three replicas logged 61, 103 and 45 slow internal subscriptions, 76, 80 and 69 slow snapshots, and 12, 11 and 15 stream leader changes.

The incident system filed a severity-3 incident at 15:50:09: "Kubernetes API unreachable" on one compute node. That node was healthy throughout. The system's own sceptic verifier, the stage that scores each candidate incident before escalation, had rated it at confidence 0.15 (as recorded in our ticket notes). The incident was escalated anyway because, by design, the verifier may suppress only incidents below severity 3. An agent closed it as misattributed at 19:17:50.

This exposes a gap that we have recorded but not yet fixed. When every probe fails with "nats: timeout", the failures share one cause, and the system should open one NATS incident rather



than a set of per-node faults.

3. Inside Ceph: Every OSD Looked Slow

At about 19:17 we read the slowest-20 op history on each SSD OSD. On all four, 20 of 20 ops exceeded 0.5 s. The slowest ops were A 2,866 to 3,385 ms (top six), B up to 4,066 ms, C 3,237 to 3,301 ms and D 2,593 to 3,117 ms. Read alone, this points at the class or at something shared.

B's ops had a different composition. On B, 12 of the top 15 were replica writes for the CephFS metadata journal and the RBD pool. Each spent 2,536 to 3,385 ms in a single stage, between started and commit_sent, which is inside the local BlueStore commit. B's slowest op (4,066 ms) was a client write to an erasure-coded pool, and 3,384 ms of it was spent waiting for shard commits. B's slow ops arrived in bursts 2 to 5 minutes apart, at 19:09:06, 19:12:07, 19:14:14 and 19:16:50.

BlueStore's own counters, averaged since OSD start, separated B more clearly. In the table below, kv_sync is the time to sync a batch of RocksDB transactions and kv_queued is time spent waiting to enter that sync. aio_wait is time spent waiting on asynchronous device I/O.

BlueStore average since OSD start, ms	A	B	C	D
kv_sync	8.80	14.73	7.96	9.69
kv_queued	8.45	32.71	7.22	13.97
aio_wait	1.85	7.06	1.52	2.61
Replica-write latency	18.75	97.21	16.67	40.44

B is worst on every row. D is second on every row, which later history explains (Section 6). The six-hour mean commit latency at 19:15 was A 19.1, B 81.9, C 14.3 and D 16.4 ms.

Several measurements looked normal. Ceph's built-in OSD bench writes 12,288,000 bytes as 4 KiB writes. It gave A 12,607, B 11,858, C 14,289 and D 12,941 IOPS, but each run lasted only 0.21 to 0.25 s. That is far too short to exercise sustained garbage collection, the drive's internal reclamation of erased blocks. The one-hour mean device write await from node-exporter was 5.1 ms for B and 2.0 to 2.4 ms for the others, and we recorded it at the time as "low".

For B we ruled out the following. SMART error counters were clean (0 reallocated, 0 uncorrectable, reserve 100). Link speed and write cache were as configured. PG and data balance were even, at 479 to 489 PGs and 18.2 to 19.6 % used. BlueStore layout matched the others. TCP retransmits were about 0.02 % on every host, so the network was not implicated. Some configuration is common to all four OSDs: BlueStore discard off, class-wide 16 MiB BlueStore throttles, and the mClock profile high_recovery_ops. A shared setting cannot by itself explain one drive diverging, though it could shape how the divergence spreads.

4. Per-Second Capture: Finding the Drive

Our first capture attempt recorded nothing because iostat was not installed on the storage hosts. The sysstat package was present with its collector disabled. We then ran iostat -dxt at one-second intervals on all four SSDs simultaneously for six minutes, alongside op history, and joined the two streams second by second.

30 Sep, 6 min at 1 s (360 samples each)	A	B	C	D
Write await p50	0.51 ms	2.06 ms	0.53 ms	0.59 ms



30 Sep, 6 min at 1 s (360 samples each)	A	B	C	D
Write await p99	9.44 ms	353.67 ms	7.23 ms	8.69 ms
Write await max	35.41 ms	725.75 ms	15.93 ms	40.39 ms
Seconds over 100 ms or 90 % util	4 (busy, under 9 ms)	17	4 (under 6 ms)	3 (under 8 ms)

The siblings' flagged seconds were high-utilisation seconds with write await under 9 ms. B's were stalls. B stalled from 19:51:01 to 19:51:06, with write await of 212.67 to 725.75 ms, 3 to 23 writes/s and utilisation of 99.9 to 100 %. It stalled again from 19:55:15 to 19:55:19, at 167.35 to 555.89 ms, 4 to 20 writes/s and 100 % utilisation. Slow ops on all four OSDs started in the same seconds: 19:51:00 to 19:51:04 and 19:55:13 to 19:55:19.

Across the second stall, A ran at 0.46 to 0.62 ms write await, 0.6 to 2.6 % utilisation and 13 to 55 writes/s. It was idle and fast, not busy. Seconds earlier, at 19:55:01, A and B had been running at the same rhythm, with B at 510 writes/s and A at 568.

The reading follows from how Ceph acknowledges writes. A replicated write completes when all replicas have committed, and an erasure-coded write completes when all shards have. The healthy drives' replica writes of near three seconds were waiting on B. The slowest member sets the latency of every write it takes part in, so a single drive's stall appears in every OSD's slow-op log.

5. The Mechanism, Re-measured

On 3 October we repeated the capture on all four drives for ten minutes and added cache-flush latency.

3 Oct, 06:31:31 to 06:41:30, 1 s (600 each)	A	B	C	D
Mean writes/s	87	74	86	71
Write await p50	0.55 ms	2.08 ms	0.53 ms	0.57 ms
Write await p99	8.3 ms	381.1 ms	6.6 ms	8.8 ms
Write await max	44.9 ms	566.9 ms	82.3 ms	51.9 ms
Seconds over 100 ms	0	15	0	0
Cache-flush await max	17.4 ms	1,072 ms	17.2 ms	17.6 ms

B had two bursts, 203 s apart. The first ran from 06:35:46 to 06:35:53 (8 s, 3 to 21 writes/s, 194 to 443 ms). The second ran from 06:39:09 to 06:39:15 (7 s, 7 to 16 writes/s, 243 to 567 ms, utilisation 100 %). Each cache flush inside a burst took 407 to 1,072 ms; outside the bursts, flushes took about 4 ms. B's median is about four times its siblings' even outside stalls, but the tail is where the damage lies.

During B's 15 stall seconds the siblings' maximum write await was A 0.8, C 6.0 and D 4.9 ms, at 31 to 34 writes/s on average. They had capacity and little to do. Their work had been held upstream, waiting for B's commit. Rates ran at 0 to 70 writes/s during the first stall. In the second after it ended (06:35:54), A, C and D completed 1,383, 1,401 and 1,392 writes and B completed 654. The queued work drained all at once.

Op history read at about 06:42 showed 14 to 18 ops over 1 s on every SSD OSD, all started within the two stall minutes. The slowest were A 4,977, B 5,263, C 5,264 and D 4,711 ms. On A, client writes spent 2,983 to 4,973 ms waiting for other OSDs' commit acknowledgements. On B,



replica writes spent 5,262 to 5,263 ms inside the local commit. The healthy OSDs waited and B worked; the second measurement shows the same pattern as the first.

6. What Aggregates Hid

Ceph exposes a per-OSD commit-latency gauge, a short-window average that our Prometheus instance scrapes every 15 s and retains for 30 days. Calendar-day means from that gauge give the history.

Day	A	B	C	D
20 to 24 Sep (range)	7.6 to 15.4	9.3 to 11.5	7.7 to 9.8	7.9 to 16.9
25 Sep (re-architecture)	22.1	88.4	13.7	111.8
26 Sep	11.8	62.6	11.8	13.0
27 Sep	11.9	89.7	10.8	12.2
28 Sep	15.5	85.2	12.4	14.4
29 Sep	17.3	82.5	13.6	15.7
30 Sep	18.6	81.1	13.9	16.1
1 Oct (B down 5.5 h, gauge 0)	14.7	48.1	12.9	14.9
2 Oct	16.0	116.5	13.4	15.8
3 Oct to 06:29	16.0	113.7	13.4	16.1

Counting 15 s samples over 1 s gives the same picture. B had 1 to 4 such samples per day from 20 to 24 September and 167 on 25 September, when D had 231. From 26 to 30 September B had 125 to 189 per day, and on 2 October it had 222. A, C and D had 0 to 4. From 26 September, B's daily maxima ranged from 3,729 to 6,389 ms. From 27 September, those of A, C and D ranged from 112 to 970 ms.

B was not historically the weak drive. From 8 to 14 September the other drives had the worse episodes: A's daily mean reached 141.1 ms and D's 96.8 ms. Two SSD OSDs other than B logged crashes in that period, one of them crash-looping. D's spike on 25 September returned to 12 to 16 ms the next day. Only B stepped up on 25 September and stayed up.

That day carried unusual write load. Metadata and erasure-coded pools were moved onto the tier by online backfill from 14:19 to 16:52. An attempted bulk move of a 1.1 TB database pool produced commit stalls up to 4.7 s and 122 slow ops in 15 minutes on B, and we redirected it to HDD. A CI scratch image of about 410 GiB (1.2 TiB raw) also sat on the tier until 30 September.

The tier mean hides most of this. B at 81 ms with siblings at 14 to 19 ms averages to about 32 ms, which looks like a modestly slow tier rather than one broken drive. Coarse per-device metrics also hide it. From node-exporter, at 15 s scrape, between 26 September and 2 October, B's daily mean write await was 3.1 to 7.6 ms against 1.6 to 2.4 ms on its siblings. Its daily maximum of the 30 s mean was 113 to 265 ms against 15 to 69 ms. We defined a stall rule at that resolution: utilisation at least 95 % with under 30 writes/s over 30 s. It matched B only 7 times in the whole period, because a stall of 6 to 8 s does not fill a 30 s window. A workable coarse signal counts 15 s steps where the 30 s mean write await exceeds 50 ms. That gives B 91 to 179 per day (27 on 1 October) and the siblings 0 to 3. Earlier node-exporter data is not comparable, because devices were renumbered on 25 September.

Short per-second samples can also miss the fault. A 90 s capture of all four drives at about 04:07 on 3 October caught no stall (B's maximum was 5.9 ms), because the stall period is 2.2 to



3.4 minutes. Cluster health stayed green between bursts. Nothing alerted on B's 25 September step for five days, because our Ceph alerts watched only health, up/in state and capacity.

7. SMART: The Drive Itself Is the Outlier

The attribute names below come from the smartmontools database entry for the WD SATA SSD family. That entry does not match our model string, so smartctl prints most of these attributes as unknown and mislabels 233 as a wear indicator. The names are therefore the vendor-family interpretation, not a confirmed mapping for this model.

Attribute (3 Oct, 06:31)	A	B	C	D
165 Block erase count	3,476	228	2,044	3,826
167 Max bad blocks per die	67	223	37	79
173 Average P/E cycles	81	133	103	49
194 Lifetime max temperature	61 °C	106 °C	64 °C	62 °C
233 NAND GB written, TLC	165,559	258,561	212,240	100,282
234 NAND GB written, SLC	34,225	2,248	20,120	37,663
241 Host writes, GiB	12,643	11,270	10,495	9,327
Reallocated / uncorrectable / reserve	0 / 0 / 100	0 / 0 / 100	0 / 0 / 100	0 / 0 / 100

The error counters that health checks usually read are identical on all four drives. The other attributes are not. B has the most bad blocks on its worst die, the highest average program/erase count and a lifetime maximum temperature of 106 °C. That maximum is undated, and current temperatures are 40 to 44 °C on all four drives. B has written 2,248 GB to its SLC cache over its life, against 20,120 to 37,663 GB on the others. B appears to send almost all of its writes straight to TLC.

Deltas over a clean interval support this. From 04:09 to 06:31 on 3 October, after the rebuild, each drive took +7 GiB of host writes. TLC NAND writes rose by +527 GB on B and +141 to +150 GB on the siblings; SLC NAND writes rose by +23 GB on B and +144 to +149 GB on the siblings. For the same host data, B wrote about 3.5 times the TLC NAND of each sibling. A longer comparison of B and A, from 19:56 on 30 September to 04:09 on 3 October, spans the rebuild. Over it, B wrote +13,721 GB TLC for +221 GiB of host writes and A wrote +3,353 GB for +174 GiB, a ratio of about 3.2.

High internal write amplification and SLC bypass are consistent with the long flush stalls in Section 5. They do not establish the cause of either.

8. The Repair and Its Regression

On 30 September we deleted the unused CI scratch image after archiving its unique data (about 75.6 GB, 903,547 files). SSD raw use fell from about 1.4 TiB to 230 GiB (3.08 %).

We then rebuilt B in place, with advisor review and operator approval. Draining B was not possible, because the k=3, m=1 pools need all four hosts, so we set noout instead. We



destroyed the OSD at 23:06:50, keeping its id, and zapped it. The operator ran a whole-device discard by hand, because our agent guard forbids automated raw-device discards. Spot reads at four offsets across the device returned zeros. We enabled BlueStore discard on B only, with one async thread, and recreated the OSD at 04:40 on 1 October. B was down for 5 h 33 min overnight while the manual step waited, and for that window the erasure-coded pools had zero redundancy margin. They kept serving. Redundancy was restored at about 04:50. Backfill finished at about 06:55: remapped PGs stood at 313 at 05:30, 298 at 06:30 and 0 at 06:55.

Post-fix, 1 Oct 05:42:20 to 05:47:20 (300 s), under backfill	A	B
Write await p50	0.52 ms	2.09 ms
Write await p99	8.45 ms	2.73 ms
Write await max	36.84 ms	2.79 ms
Writes/s	140	450

B had zero seconds over 50 ms in that window. Afterwards the slowest ops on the four OSDs were 393 to 950 ms, none over 1 s, where previously 20 of 20 had been near 3 s.

The improvement did not last. B's hourly mean commit latency was 16.6 to 19.7 ms in the hours ending 08:00 to 14:00 on 1 October, with one hour at 35.6. It then rose to 59.3 at 15:00, 101.6 at 17:00 and 142.0 at 23:00. From then until 06:00 on 3 October it ranged from 38.8 to 325.6 ms, mostly 60 to 165, while the siblings stayed at 12 to 19 ms. The fix held about ten hours after the recreate and about eight after backfill ended.

NATS followed the drive. Its warnings ran at 0 per hour from 00:00 to 05:00 on 1 October while B was destroyed, then 0 to 7 per hour through 13:00. They rose to 4 to 19 per hour from 14:00 to 23:00, and reached 83 in one hour on 2 October.

Day (three replicas)	Slow subscriptions	Slow snapshots
29 Sep	103	116
30 Sep	98	108
1 Oct	74	41
2 Oct	237	156
3 Oct to 06:30	83	62

The longest snapshot on each day took 7.70 to 9.56 s.

A fifteen-minute capture of B and one sibling on 3 October showed the regressed state.

3 Oct, 04:09:24 to 04:24:23, 1 s (900 each)	A	B
Mean writes/s	84	90
Write await p50 / p99 / max	0.53 / 9.0 / 24.5 ms	2.06 / 404.2 / 623.5 ms
Seconds over 100 ms	0	41
Seconds over 100 ms or util over 95 %	2	47
Cache-flush await max	26 ms	1,012 ms

B had six bursts of 6 to 7 s each, at 04:09:45, 04:12:14, 04:15:02, 04:18:09, 04:20:21 and 04:22:44, with gaps of 132 to 187 s. During them it ran at 2 to 44 writes/s and 142 to 624 ms write await, with queue depth up to 18.4 and flushes of 278 to 1,012 ms each.

The newly enabled discards are not the trigger. B issued discards in 800 of the 900 seconds, at a mean of 25.5/s and about 1 ms each outside stalls, but at most 3 in any one second inside a



burst. Per-minute Prometheus data for the preceding three hours gives B a p50 of 15, a p99 of 3,079 and a maximum of 3,401 ms, with 11 samples over 500 ms; A's maximum was 84 ms. B now holds 65 GiB (3.51 % of its capacity) on a device fully discarded on the night of 30 September. A shortage of free blocks cannot explain the regression.

9. Relation to PT-R-2026-031

PT-R-2026-031, "The Pool Named NVMe Lives on Spinning Disk" (1 October), covers the same tier from the other side. Consumer-SSD slow ops on 25 September led us to put the database pool on HDD. On 27 September we fixed the stalled git forge by moving its 536 MB SQLite database onto the replicated SSD RBD pool, which is this tier. Nothing here contradicts that paper, but several of its observations now have an owner.

The two SSD OSDs in 031's 25 September episode were B and D. That paper called D the weakest that morning (aio wait 13.7 ms), which was true that day, but D recovered from 26 September and B did not. The Phase B "commit stalls up to 4.7 s, 122 slow ops in 15 minutes" described in 031 were on B. Paper 031 left two SSD-tier spikes on 27 September unexplained: 3,472 ms at 06:50 and 2,826 ms at 07:20. Per-OSD queries attribute both to B. Over the 10 minutes ending 06:50, B's maximum was 2,912 ms against 76 to 148 ms on the others; over the 10 minutes ending 07:20 it was 3,289 against 41 to 102 ms. The peaks differ from 031's because the windows differ, but the attribution does not. That paper's figure of "SSD 12 ms vs HDD 82 ms average commit" was a four-OSD mean, the kind of aggregate shown in Section 6 to hide one slow member. Both figures stand as means, and so does the forge's improvement.

The forge is now exposed to B. Between 02:24 and 05:16 on 3 October it logged 19 slow-SQL warnings of 5.21 to 7.66 s at 8 timestamps. These covered single-row runner, key and token lookups and a runner heartbeat update. At 5 of those timestamps B's commit gauge peaked at 203 to 525 ms within 90 s, while A, C and D stayed at or below 115 ms. This is consistent with B's stalls but not proven event by event. It does not reverse 031: on HDD under co-tenant load the forge was unusable. It does mean that the SSD tier's light-workload role now carries a multi-second tail set by one drive. Paper 031's lesson, never to bulk-migrate onto the consumer SSDs, agrees with B's step on the day of that migration. Whether the rewrite damaged B or exposed a drive that was already weak is not established.

10. What These Numbers Will Not Carry

The population is four drives of one model in one cluster, and the finding concerns one of them. Nothing here supports a claim about the WD Red SA500 in general, or about consumer SSDs as a class.

The per-second captures are short. They cover 6 minutes on all four drives, 5 minutes post-fix on two drives, 15 minutes on two drives and 10 minutes on all four. The stall periodicity rests on six bursts in one capture and two in another. The raw iostat files from 30 September were not retained, so those figures come from the analysis output recorded in the session; the 3 October files are retained. The post-fix measurement was taken during backfill, not at steady state, and it is the result that later regressed. It shows that B could run cleanly after a full discard; it does not show that the rebuild repaired anything durable.

The Ceph commit gauge is a short-window per-OSD average sampled every 15 s. It sometimes reads 0 while the OSD is up (from 02:20:00 to 02:25:30 on 3 October), and it read 0 throughout B's outage, so B's 1 October mean is understated. The SMART names are a vendor-family interpretation. Host writes have 1 GiB resolution, so the +7 GiB interval carries about 14 % quantisation, which bounds the precision of the 3.5 times ratio. The 0.15 verifier confidence comes from our ticket notes and was not re-read from the raw verdict.



There are confounds in the timeline. All OSDs restarted for package upgrades on 2 October, with B restarting at 22:58. B also reported down for about a minute at 20:44 and at 22:32 on 1 October. The regression was visible from 15:00 on 1 October, before any of these events, so they do not explain its onset. They may still affect later figures.

Several figures in our own working summaries were wrong, for example a peer replica-write range, a stall count that mixed definitions, and the duration the rebuild held. This paper uses the raw counters throughout.

Some claims survive these limits. B is the outlier on per-second write and flush latency in every capture. The tier-wide slow ops coincide second by second with B's stalls and show the healthy OSDs waiting on B. B stepped up on 25 September and stayed degraded while its siblings did not. B's SLC usage and NAND write ratio differ from its siblings' by large margins. The rebuild's benefit lasted about ten hours.

Other claims do not survive. We cannot say why B degraded: a defect, the 25 September rewrite, the undated 106 °C, or firmware behaviour on this unit all remain open, and there is no vendor diagnosis. We cannot say why the rebuild held as long as it did. Clean blocks after the discard are plausible, but we did not measure them. We cannot say whether the siblings will follow, whether discard would help them, or whether a replacement drive will remove the tail. The forge's slow queries are linked to B only by coincidence in time.

11. What We Changed and What Comes Next

We added a flash-OSD commit-latency alert that fires when the six-hour mean exceeds 40 ms and holds for 6 h. It was tested to fire on the 25 September shape and to stay silent on HDD OSDs and on a one-hour 150 ms burst. A negative control with a 200 ms threshold fails that test. The alert went pending at 19:59:42 on 30 September, with B's six-hour mean at about 76 ms. It fired from 02:00 to 02:50 on 1 October, went pending again at 16:36:57, and has fired since about 22:40 on 1 October. We have left it on because it is accurate. Our per-second capture tooling is now known to depend on iostat being present, which it was not.

The drive has not been replaced. The intended replacement, an enterprise SATA SSD with PLP, has not been bought or fitted; it needs a purchase and an on-site visit, which is an operator decision. Remote levers can only soften the tail. B cannot be drained without losing erasure-coded redundancy, and setting its primary affinity to zero would help reads but not the stalled replica writes. The next experiment is the replacement itself: the same all-four-drive per-second capture before and after, run for longer than ten minutes. The incident system's per-node attribution of bus-wide NATS failures remains unchanged and is the other open item.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. This paper is PT-R-2026-035. The measurements were taken between 30 September and 3 October 2026 (all times UTC, converted from BST where the raw timestamps used it), with Prometheus history from 3 September to 3 October. Raw artefacts are retained, except the 30 September iostat files noted in Section 10.