SEVERITY 1 · STORAGE SUBSYSTEM · FOLLOW-UP

Storage Array Multi-Device Drop: Recovery Follow-Up

Incident WI-2026-0731-01 · Follow-up report · 7 August 2026

ENVIRONMENT	PERIOD COVERED	SEVERITY	STATUS
Internal Provider Infrastructure	1 to 7 August 2026	1 · Critical	Full redundancy restored

Executive summary

SUMMARY

The storage array that lost four of its twelve drives on 31 July is now back to full redundancy. It has all twelve drives, tolerates two failures again, and the array itself reports no errors.

Getting there took six days, and that was deliberate. The array came back on 31 July running on ten drives with no capacity to absorb another failure. Repairing it means several hours of the heaviest work the machine ever does, on a machine whose fault was not yet understood. Walter Infrastructure (WI) copied the data off first and repaired second, so that a failure during the repair would be recoverable. The copies finished on 3 August and were verified. The repair began the same morning.

One drive failed during that repair, on 3 August. It was not a surprise in kind, since the array was operating without redundancy, and the copies were already complete. That drive was found to have physically damaged areas on its surface. It was replaced on 6 August and the array rebuilt overnight to full redundancy.

Executive summary, continued

SUMMARY

Two findings from this period change the conclusions of the first report.

The first report identified the failure mechanism but left the trigger unexplained. The trigger is now identified. The appliance runs a scheduled routine every morning that reads health data from every drive at once, and the controller fault reproduces immediately behind it, on schedule. The fault has been observed on five separate mornings. On none of those occasions did it remove a drive from the array or produce a filesystem error.

The second finding concerns what is at fault. When the failed drive was removed, the daily errors stopped for thirteen hours, which suggested that one bad drive had been provoking the shared controller. That reading was tested and did not survive: the errors resumed on schedule with the drive physically out of the machine, affecting all remaining drives. The shared controller stage is therefore weak independently of any one drive. This is the finding that determines the replacement hardware requirement.

MEASURE	STATUS
Array redundancy	12 of 12, full dual parity
Array state	Clean
Data loss	None found
Volume integrity	Verified, all datasets present and readable
Filesystem errors, whole event	Zero
Devices replaced	1
Failure mechanism	Identified
Trigger	Identified, reproducible
Controller fault	Recurring, non-destructive, monitored

> This report covers WI's own internal environment, not a client system, and is published as a worked example of the incident process. It is redacted for distribution.

1. Data protection performed before repair

METHOD

22 TB was copied to two separate appliances and verified before any repair action was taken. The sequence was ordered by reacquirability, highest value first, so that the least replaceable data was protected earliest.

DATASET	SHAPE	COMPLETED	VERIFICATION
A	Single sequential object, 1.5 TB	1 Aug 05:38	Byte count identical both sides
B	Many-object, 26,904 objects, 1.18 TB	1 Aug 08:52	26,904 of 26,904 present
C	Many-object, 8.4 TB	2 Aug 06:39	Object-level comparison, complete
D	Many-object, 12.1 TB	3 Aug 06:49	Object-level comparison, complete

The run took 54 hours and recorded no transfer errors, no retries and no stalled mounts.

Datasets A and B were treated as the protected set and completed first, on the reasoning that they have no reacquisition path. Datasets C and D can be reconstructed from external sources and were copied second.

2. Verification method, and a correction it produced

METHOD

Copy completion was verified by comparing object inventories between source and destination. Size totals were not used, and that mattered here.

For Dataset D the totals were misleading in both directions at once. The destination held more objects than the source, because the copy ran without a delete pass and retained superseded versions of objects that had been replaced during the run. The destination also held fewer bytes than the source, because objects added late in the run had not been captured. The two effects partly canceled, producing a total that looked approximately correct while thirteen objects were in fact missing.

The inventory comparison found them. They were copied in a follow-up pass and the comparison re-run, returning zero missing of 15,776.

A size comparison would have reported this copy as complete. Size totals are not a verification method for a dataset that is being written to during the copy.

3. Pre-repair audit, and content that was not in scope

FINDING

Before the repair, the volume was audited against the copy set to confirm nothing had been missed. Three datasets were found that had never been in scope, totaling 8.5 GB.

DATASET	SHAPE	REACQUIRABLE
E	179 objects, 5.2 GB, authored work product with an associated project file	No
F	20 objects, 12 MB, no external source	No
G	3.3 GB, curated collection assembled by hand	Not practically

The rescue plan had been scoped around a stated understanding of what could not be replaced. That understanding was incomplete. Dataset E in particular is authored work with no external source, and it had been on the array long enough that its presence was not recalled.

All three were copied and verified before the repair proceeded. The copies were correct against the plan. The plan was wrong. What found the gap was auditing the volume itself, and that ordering is now a standing step.

4. Repair execution

METHOD

The appliance management interface could not perform the repair. Its rebuild dialog reported zero devices available out of sixteen and would not proceed.

The cause was the stale device error flags described in Section 7 of the previous report. Those flags mark a device unhealthy in the appliance database, and the management interface will not offer a device it believes is unhealthy as a rebuild target. Both candidate devices carried flags set during the original fault. The array layer showed both as available, correctly sized and error-free, and the array itself as clean and degraded with zero failed devices.

The repair was performed at the array layer instead, which added both devices online without stopping the array or interrupting service.

Both devices were added within seconds of each other, deliberately. A dual-parity array reconstructs two members in a single pass if both are present before reconstruction begins. Adding the second afterwards queues an independent second pass, roughly doubling both elapsed time and the period of sustained load. Measured throughput was 72 MB/s reconstructing one member against 49 MB/s reconstructing two, so a single pass costs materially less time and sustained load than two sequential passes.

5. Device failure during reconstruction

EVENT

Approximately forty minutes into reconstruction, one of the two devices being rebuilt failed and was removed from the array by the kernel.

The failure signature was distinct from the original event. The device accepted 299 queued commands and completed none of them. The controller aborted the queue, issued a hard reset, and the device did not return. The kernel marked it offline after seventeen minutes of recovery attempts.

Scope was contained. Each of the following was measured:

- Zero filesystem errors. No read-only remount, no journal abort.
- All 27 I/O error entries named the failed device and no other.
- The volume remained mounted read-write with all data present and readable.
- The remaining eleven devices passed direct read tests.
- Reconstruction of the second device continued without interruption and completed successfully.

The array reached eleven of twelve members and single-device fault tolerance was restored.

6. Failed device analysis

ANALYSIS

The failure was caused by damaged areas on the drive surface. The evidence points to the device itself and away from its connection.

The I/O errors cluster on specific physical regions. Twenty-six of twenty-seven fall within a 3.4 MB band approximately 125 GB into the device. The twenty-seventh falls within 36 MB of an unrecoverable read error the same device had logged two days earlier, at a different location. That indicates at least two separate damaged regions.

Interface error counters on the device read zero, identical to a healthy peer. Signal integrity faults in cabling, connectors or backplane increment those counters. They did not move, which points away from the connection and to the device.

The behavior is consistent with a drive entering internal recovery on marginal media, ceasing to service its command queue while it retries, and exceeding the controller timeout. A drive that returns an error is behaving correctly. This device accepted hundreds of commands and then stopped answering.

> **Why it surfaced during reconstruction and not before:** ordinary use does not read those regions. Parity reconstruction reads every sector on every member, so it reached areas that ordinary use had not touched in years. Reconstruction is the first workload that tests a whole device, which is why a device failing during reconstruction is common and why copies are taken first.

7. Device replacement and return to full redundancy

METHOD

The failed device was replaced on 6 August. The old device was removed from the array cleanly before physical removal so that it left no stale entry.

The appliance handled the replacement device without intervention: it detected it, wrote the full partition layout, and added it to the array automatically. A device carrying no error flag proceeds through the normal automated path, which confirms the diagnosis in Section 4. The flags were what blocked the repair.

Reconstruction ran overnight and completed. The array reports twelve of twelve members, state clean, zero failed devices.

The replacement device has no readable service history. It reports zero power-on hours, zero power cycles and one start-stop event. Those counters have been reset. The device therefore has no readable service history, and its accumulated wear is unknown. It is monitored on change from its installation baseline. That detects developing faults but cannot establish remaining life. It came through a full reconstruction, the heaviest workload it will encounter, with all error counters still at zero.

8. Trigger analysis: resolved

ANALYSIS

The previous report recorded the trigger as unexplained. It is now identified.

The appliance runs a scheduled routine each morning that collects health telemetry from every device. The controller fault reproduces within minutes of that routine, on schedule, and has been observed doing so on five separate mornings.

This is consistent with the failure mechanism described previously. The fault appears under concurrent access to all devices behind the shared controller stage, and the telemetry sweep is the only routine workload that touches every device at once.

> **The fault is reproducible on demand.** That is a materially better position than an unexplained intermittent fault. It also means the condition can be evaluated on the replacement hardware before committing to it.

Observed consequences across all occurrences: no device removed from the array, no filesystem error, no data affected. The signature is treated as known and non-destructive, and is monitored without alarming.

9. A hypothesis raised, tested and rejected

ANALYSIS

When the failed device was removed on 3 August, the controller errors stopped completely for thirteen hours, including the eight and a half hours of parity reconstruction that followed. Reconstruction is the heaviest concurrent load the controller sees. Passing it in silence was strong apparent evidence that the failing device had been the source of the fault all along.

That reading was recorded as provisional and identified the daily telemetry sweep as the test that would settle it.

It did not survive. The errors resumed on the next scheduled sweep, with the device physically removed from the machine, and affected all eleven remaining devices. The thirteen hours of quiet were simply the interval between one daily sweep and the next.

HYPOTHESIS	VERDICT
The failing device was provoking the shared controller stage	Rejected by observation. Fault recurred on schedule with the device removed.
The shared controller stage is weak independently of any one device	Supported. Fault occurs across all devices with the suspect device absent.
The failed device had damaged media	Supported by measurement. Errors cluster on two physical regions; interface error counters at zero.

The rejected reading was believed and acted upon for thirteen hours, and the conclusion it would have overturned governs replacement hardware selection. A single quiet interval is not a controlled test when the fault has a daily period.

10. Stale device error flags: cleared

CORRECTION

The previous report described four stale device error flags left by the original fault, marking devices as failed that were demonstrably healthy, and recorded that clearing them was expected to require a service interruption.

That was incorrect. **The appliance provides a supported command that clears a device error state directly. All four flags were cleared while the array was mid-reconstruction. The appliance's own health view moved from four devices reporting abnormal to sixteen reporting normal, the front panel indicators cleared, and the array was undisturbed throughout.**

The flags matter more than their cosmetic effect suggested. A single stored flag drives the panel indicator, the health field in the management API, the device status in the array member list, **and the exclusion of that device from the rebuild dialog**. The inability to repair through the management interface, described in Section 4, was caused by these flags.

They do not clear themselves. A flag is written when an error occurs and is removed only by an explicit clear. A device that returns to normal operation never generates one, so a flag set during an incident persists indefinitely.

One flag remains and is retained deliberately. It is keyed to the identifier of the failed device, which is no longer in the machine, so nothing present matches it. If that device were reinstalled it would be flagged immediately, which is correct.

11. Device health telemetry: gap closed

CORRECTION

The previous report recorded that per-device health data could not be obtained on this platform, and that device health was therefore inferred from performance.

That gap is closed. Full per-device attribute tables are available on the appliance, refreshed daily by the scheduled telemetry routine. A summary is also available through the management API. The standard command-line tooling is absent, which is what produced the original conclusion, but the data itself was present throughout.

One caution applies to the API route. Its health field reports stored fault state, not current condition: it reported four devices abnormal while two of those four were healthy, in-sync array members with all attribute counters at zero. The attribute tables are authoritative. The health field is not.

A second caution applies to comparison across vendors. The replacement device is from a different manufacturer than the eleven originals and encodes several attributes as packed rate values rather than counts. Read literally against the other devices' zeros, three of its attributes appear to show large error counts and do not. Device health on a mixed-vendor array must be assessed per vendor.

Health baselines were captured for all twelve devices and are now the reference for change detection.

12. Current state

STATUS

ITEM	STATE
Array redundancy	12 of 12
Array state	Clean, zero failed devices
Volume	Mounted read-write, all datasets present
Filesystem errors since original event	Zero
Devices reporting abnormal	Zero
Panel fault indicators	Clear
Daily controller fault	Recurring, no member loss, monitored
Protected datasets	Second copies retained on two separate appliances
Services	All restored

Second copies of the protected datasets are retained. They are not a permanent storage tier and their disposition is already scheduled.

13. Actions completed

STATUS

ACTION	COMPLETED
22 TB copied to two appliances, verified by inventory comparison	3 Aug
Pre-repair audit; three additional datasets found and copied	3 Aug
Array repaired at the array layer after management interface failed	3 Aug
Failed device analyzed; damaged media identified	3 Aug
Failed device replaced	6 Aug
Array reconstructed to full redundancy	7 Aug
Trigger identified and reproduced	5, 6, 7 Aug
Device-as-trigger hypothesis tested and rejected	5 Aug
Four stale device error flags cleared without service interruption	6 Aug
Per-device health telemetry route established; baselines captured	6 Aug
Backup platform retention corrected; 881 GiB reclaimed	1 Aug

14. Pending actions

STATUS

ACTION	STATUS
Observation period under normal load, approximately one week	In progress
Replacement enclosure selection	Specified, candidates evaluated
Disposition of retained second copies	Scheduled, follows enclosure replacement
Backup coverage gaps identified during this work	Tracked separately

The observation period's pass criterion is device loss from the array. Controller log activity alone does not fail it. The daily controller fault is expected and has been shown across five occurrences not to affect array membership or data. A device leaving the array would be a different finding.

15. Replacement hardware requirement

ASSESSMENT

The finding in Section 9 sets the requirement directly.

The fault is in a stage shared by all twelve devices. It is not attributable to any individual device, and it reproduces on demand under concurrent all-device access. Replacement hardware must therefore give each device an independent path to the host, rather than multiplexing devices behind a shared stage.

This is a specific, testable requirement. It can be verified against a candidate before purchase by reproducing the all-device access pattern.

Most appliance vendors do not publish how their bays are wired. For a twelve-bay platform it is a question that always has an answer, because no mainstream processor in that class provides twelve native ports. Something is in the path. Where topology could not be established from vendor documentation, the candidate was not considered.

16. Assessment

ASSESSMENT

The array is fully redundant, clean, and has lost no data. The failure mechanism was identified in the first report and the trigger is identified in this one. The fault is reproducible, which makes the replacement decision testable.

The six days at reduced redundancy came from a deliberate ordering. Copying 22 TB and verifying it took 54 hours, and repairing before that completed would have meant performing the heaviest workload the machine does, on an array with no fault tolerance, without a second copy of the data. One device did fail during the repair. Had that repair run first, the same failure would have occurred against unprotected data.

Two corrections to the previous report are carried here. The stale device flags did not require a service interruption to clear, and per-device health telemetry was available throughout. Both were stated at the width of what had been checked at the time, and both checks were narrower than the conclusions drawn from them.

The appliance is treated as serviceable but not proven. It carries a known, reproducible fault in a shared component that has not caused data loss across five observed occurrences and one full reconstruction. Replacement is specified and in progress. The protected datasets hold verified second copies in the meantime.

Walter Infrastructure
Incident WI-2026-0731-01 · Follow-up · 7 August 2026