Failure Testing Report

Example Deliverable — 4-Node Hyperconverged Cluster

Document	**Example deliverable (sanitized)**
Prepared by	**Joy Automation**
Version	1.0
Date	July 31, 2026

Sanitized example from a completed engagement — client-identifying details removed. Provided to illustrate the deliverables Joy Automation produces.

Contents

> *Sanitized example deliverable from a completed Joy Automation engagement, included to illustrate the failure-testing documentation you would receive with a delivered system. It reports results for that project's 4-node, two-site cluster; your delivered report will reflect tests run against your own as-built architecture. Client-identifying details have been removed.*

Joy Automation · failure-injection campaign, 2026-06-08 → 2026-06-09 · all five tiers executed.

This report documents **what we tested**, **what we fixed as a result**, and **what each failure mode means for operations** (the operator procedures live in the Operations Guide §7 Runbooks — this report references them rather than repeating steps).

Companion document: the Operations Guide — day-2 ops + the recovery runbooks referenced here. Full per-test evidence (raw 50 Hz probe logs and per-tier result notes) is retained by Joy Automation and available on request.

---

1. Method

Every claim was **measured by injecting a real fault and timing the recovery**, against a verified-green baseline (Ceph `HEALTH_OK` / 545 active+clean, k8s 4× Ready, CNPG 3/3, ring Idle) restored between tests. Faults were injected with:

- **Hard node kill:** `echo b > /proc/sysrq-trigger` (instant kernel reset — worst case, no clean shutdown).
- **Network/ring:** at the switch CLI (`ssh admin@172.16.150.4N`), `configure terminal → interface … → shutdown` (real switch-port down).
- **Service/component:** native controls (`ceph mgr fail`, `incus move`, `kubectl delete pod`, `ceph osd out`).
- **Probes:** a 50 Hz availability-sampling probe for RTO timing.

Safety: a BMC power-cycle backstop was on hand for the catastrophic tier (never needed — all nodes self-rebooted in ~3 min). Restore drills used throwaway targets or live systems at the owner's direction, and never destroyed production data.

---

2. What we tested (results by tier)

Tier 1 — Component / service failover (non-disruptive) · all PASS

Test	Injection	RTO	Verdict
T1.1 CNPG primary failover	delete `tsdb1-1` pod	**~3 s** write-gap	✓
T1.2 MetalLB L2 failover	delete announcing speaker	**<0.3 s** (below probe resolution)	✓
T1.3 Ceph mgr failover	`ceph mgr fail`	dashboard followed to standby	✓
T1.4 RGW ingress failover	stop one `rgw` daemon	**0** (VIP served throughout)	✓
T1.5 Incus stateful live-migration	`incus move ignition` N1→N3→N1	**~1.4 s** cutover blip	✓
T1.6 ceph-csi RBD reschedule	(via T1.1)	no `Multi-Attach` stall	✓
T1.7 K8s API VIP failover	reboot VIP holder	**0** `/healthz` failures	✓

Tier 2 — Single hardware failure (hard kill, one node at a time) · all PASS

Test	Injection	Key result	Verdict
T2.1 Worker node (node4)	sysrq reset	Ceph PGs stayed active (min_size=2); API unaffected; clean ~51 s post-return	✓
T2.2 Control-plane node (node3)	sysrq reset	mon quorum 2/3 held; etcd writable; **API VIP ~22 s**; clean ~23 s post-return	✓
T2.3 ★ Heavy node (node1) + healing eval	sysrq reset	Incus auto-heals VMs onto survivors (~40 s, same IP); decision gate for VM-HA design	✓
T2.4 OSD/disk failure + replace	`osd out` / `--replace`	non-event; backfill ~500 MiB/s; clean ~24–30 s	✓
T2.5 LACP bond member loss	down one bond member	**0 packet loss**	✓

Tier 3 — Network / ring / switch (ERPS) · all PASS

Test	Injection	Result	Verdict
T3.1 Ring-link, intra-site (L1)	shut sw1 tg0/27	ERPS Idle→Protection; **0.05 % loss ≈ <20 ms**; all nodes reachable; revertive	✓
T3.2 Ring-link, inter-site (L2)	shut sw2 tg0/27	reconverged **~220 ms**; ring held	✓
T3.3 Switch loss	(one switch down)	that switch's node drops (→ behaves like T2.1); ring re-forms around it	✓

Tier 4 — Catastrophic (site loss + cold start)

Test	Injection	Result	Verdict
T4.1 Site-B loss (node3+4)	both down	**FULL OUTAGE** of k8s + DB; manual recovery ~5–15 min; no data loss	⚠ by design
T4.2 Site-A loss (node1+2)	both down	**FULL OUTAGE**; surfaced the *apiserver-stuck-on-VIP* failure mode (kubelet restart)	⚠ by design
T4.3 Full cold start	all 4 powered on from off	auto bring-up; ~4 min to API, ~10 min full green	✓

> **Headline:** the cluster cleanly survives **any single node**, but **no two-node / full-site loss** — that's a quorum property of 2 sites, not a defect: Ceph mons, etcd, and Incus dqlite each need a *majority* of voters, and two equal sites can't form one when half is gone. Removing the limit needs a **3rd failure domain** (a small tiebreaker box/VM in a separate location — quorum traffic only, not data) plus Ceph stretch mode (`size=4`, 2 copies/site, `min_size=2`, 5th tiebreaker mon, ~25 % capacity cost), a 3rd-site etcd member, and a 3rd-site Incus voter. Documented as a forward path, not a defect to tune away.

Tier 5 — Restore drills · PASS (mechanism)

Test	What	Result	Verdict
T5.1 CNPG/TimescaleDB restore	recover a throwaway DB from `s3://cnpg-backups`	restored from backup alone (~100 s); backups valid	✓ mechanism
T5.2 Ignition `.gwbk` backup/ restore	backup → change → restore on live gateway	3.67 MB `.gwbk`, restore overwrote+restarted cleanly	✓ mechanism

Caveat: both restores were validated with little/no production data. **Re-validate fidelity once the historian and Ignition carry real data** (§4).

3. What we mitigated (found and fixed during testing)

#	Discovered	Root cause	Fix	Where
1	MetalLB LB VIP failover stalled ~4 min on worker kill	CP nodes carried Talos' `node.kubernetes.io/exclude-from-external-load-balancers` label → only the single worker announced	Patch removes the label so all 4 nodes announce; failover now **~6 s**	`k8s/talos/lb-announce-cp.yaml`
2	Windows VM "NIC passes no traffic"	**Misdiagnosis** — NIC fine; Windows Firewall (Public) dropped ICMP	`create-winserver` adds an inbound ICMP-allow rule	`create-winserver`
3	VMs failed to auto-start after full power loss	`incus-startup` raced ahead of Ceph **and** of dqlite cluster formation	systemd drop-in waits for both RBD **and** `incus list` before starting VMs	`incus-startup-wait-for-ceph` drop-in
4	A lone-survivor control-plane node held the API VIP but the apiserver was wedged	apiserver stuck after quorum loss/recovery	`talosctl service kubelet restart` on that CP node	Ops Guide §7.7
5	VM auto-recovery design (fencing daemon vs native)	risk of split-brain double-write on a "partitioned but alive" node	Chose **native Incus healing + RBD exclusive-lock** (storage-level fence) + rolling snapshots — no custom daemon to own (a bespoke IPMI-fencing daemon's false-positive mode, powering off a *healthy* node, can be worse than the outage it prevents)	design decision
6	Ignition gateway documented as Docker	actually a native systemd install (`/usr/local/bin/ignition/`)	Corrected docs; drove backup via `gwcmd.sh`	`gwcmd` / Ignition

Resilience-tuning applied as a result: `cluster.healing_threshold` lowered (not aggressively), `cluster.evacuate=live-migrate` + `migration.stateful=true` for ~1.5 s planned-maintenance cutover, and RBD exclusive-lock + `vm-snapshots` enabled cluster-wide.

4. What this means operationally (discovered → runbooks & follow-ups)

Each failure mode maps to an operator procedure in the **Operations Guide §7**:

Failure mode (tested)	Operator action	Runbook
Single node dies	Verify green; `incus cluster restore` when it returns (self-reboots reliably)	§7.1
OSD / disk fails	Replace drive → `ceph orch osd rm --replace`	§7.4
Switch / ring / link fails	Verify ERPS re-formed; restore hardware (auto WTR)	§7.5
K8s API / LB VIP wedged	`kubelet restart` (API) / check MetalLB speakers (LB)	§7.6
Two-node / full-site loss	**Manual recovery (~5–15 min)** — self-reboot/BMC, `cluster restore`, kubelet check	§7.7
Full power loss	Cold-start sequence (~4–10 min to green)	§7.8
Data loss / bad change	CNPG restore from S3 · Ignition `.gwbk` · VM snapshot rollback	§7.3 / §7.9

Operational follow-ups discovered by the campaign (also in Ops Guide §9–10):

1. **Keep BMC/IPMI creds accessible** — node self-reboot was 100 % reliable, but a *hung* node needs a remote power-cycle.
2. **Re-validate restore fidelity** with real historian/Ignition data post-commission.
3. **Plan the CNPG Barman-Cloud-Plugin migration** before any CNPG operator 1.30 upgrade (current native Barman is deprecated upstream).
4. **Wire plant monitoring/alerting** (none external today) — alert on mon-quorum loss, OSD near-full, NVMe temp, ERPS state.
5. **If full-site survival is later required**, scope the **3rd-site tiebreaker + Ceph stretch mode** (`size=4` / 5th tiebreaker mon + 3rd-site etcd & Incus voters — see the Tier 4 headline above).

5. Survivability at a glance

Failure	Outcome	RTO
Any single node / disk / NIC / switch / link	**non-disruptive** (auto-recover)	0 s – ~40 s
Planned node maintenance	live-migrate	**~1.5 s**
Database primary / API node / LB VIP	brief auto-failover	~3 s / ~22 s / ~6 s
Two nodes / a whole site	**full outage, manual recovery, no data loss**	**~5–15 min**
Full cold start	auto bring-up + verify	**~4–10 min**

Verdict: single-fault tolerance — the design standard — is **met and proven**. The two-site footprint cannot survive a full-site loss without a third failure domain; that is documented as a forward path, not an open defect.

6. Evidence index

Backing evidence for this report — retained by Joy Automation, available on request:

- Per-test detail & raw timings (per-tier result notes + 50 Hz probe logs).
- The test specification/catalog and campaign field notes & decisions.

The as-built system state at delivery — switch running-configs and per-node network configuration — is included in your handover package (snapshot on 2026-06-09, with a final pre-ship update on 2026-06-11).