Cluster Storage Vendor Assessment Workflow

Field	Value
Author	Eva Winterschön
Section	research/vendor-assessment-workflow
Version	0.2.0
Date	2025-08-08
Repo	https://github.com/evaw-cerebras/
Summarized	HPC, Cluster Storage, Performance Benchmarking
Aggregates	Docs + Reqs + Components (June+July 2025), RAG Analysis
Inferenced	Qwen3-235B-Instruct

0 | Governance & Preparation (1 week)

Activity	Outcome
Appoint evaluation board (technical lead, procurement, InfoSec, finance)	Named RACI matrix
Publish charter & acceptance criteria (“decision must be data-driven; no single vendor veto”)	Signed charter in repo
Stand up collaboration space (Git/SharePoint/Jira)	Single source of truth

Exit gate: Charter approved, budget/time boxed.

1 | Requirements & Workload Characterization (2 weeks)

Workload taxonomy

- Mixed read-heavy AI inference, small-file checkpointing, parallel writes (≥ 128 KiB), random metadata ops.
- Peak client fan-out: ≥ 32 compute nodes × (2 × 200GbE Dual-Port or 2x 400GbE Single-Port RoCEv2 NICs)
- Capacity growth: 1 PB → 5 PB in 36 months; ≤ 3 ms 99p latency on 4 KiB random reads.

Non-functional needs

- Protocols: NFS v4.1 over RDMA, S3 (SigV4)
- RHEL 8, 9 compatibility: validated kernel modules, OFED drivers
- Security: FIPS-140-3 crypto, AES-256 at rest, KMIP key rotation (reference CISO checklist)
- Reliability: ≥ 5 years media endurance (≥ 1 DWPD TLC / ≥ 0.2 DWPD QLC), N+2 node HA, rebuild < 4 h for 24 TB NVMe.
- Manageability: fully documented REST/Ansible collection; SNMP v3 / Redfish telemetry.

Success metrics (placeholders)

Metric	Pass / Stretch
4 KiB random read 99p latency	≤ 3 ms / ≤ 2 ms
256 KiB sequential read throughput (8 clients)	≥ 140 GB/s / ≥ 160 GB/s
1 MiB PUT S3 throughput (16 threads)	≥ 20 GB/s / ≥ 25 GB/s
Fail-over recovery time	≤ 60 s / ≤ 30 s

2 | Market Scan & Long-List (1 week)

- Sources: SNIA vendor lists, IDC market-share, recent SPC-1/SPECsfs/NVMe-oF results.
- Create vendor inventory sheet capturing: product family, controller CPU, media type (TLC/QLC), scale-out limits, software licensing, public reference sites.

Exit gate: Top 5–7 vendors invited to RFI.

3 | RFI / Paper Evaluation (3 weeks)

Structured RFI template

Spreadsheet with yes/no & mandatory commentary, covering ~150 items across:

- Protocol compliance (NFS4.1 + pNFS, object versioning, IAM integration)
- Data-services (snapshots, erasure coding, QoS/per-filesystem caps)
- Sustainability (W/IOPS, refrigerant-free cooling)
- Road-map commitments (PCIe 5.0, CXL storage class memory, 800 GbE)

Scoring Methodology

- Numeric: (0 = miss, 1 = partial, 2 = meets, 3 = exceeds);
- Weights: technical 50 %, commercial 20 %, support 15 %, ESG 15 %.
- Drop vendors scoring < 65 % or failing any Mandatory.

Exit gate: Short-list (≤ 3) for PoC.

4 | Laboratory Proof-of-Concept (6 weeks)

4.1 Test-bed

Component	Spec
Compute clients	12 × RHEL 9 + KVU / KVSS stack
Load-Gen clients	3 × RHEL 9 + load-generating stack
Per-Node Fabric	4x Thor2 1x400GbE RoCEv2, leaf/spine
Switch telemetry	In-band-network-telemetry (INT) enabled
Test tools	**fio**, **vdbench**, **IOR/mdtest**, **cosbench/s3-bench**

4.2 Benchmark Plan

Test	Why
Synthetic micro-bench (fio 4 KiB/4 KiB, 70/30 RW)	isolate device latency
Mixed I/O profile (70 % 128 KiB seq-read, 30 % 4 KiB rand-write)	mimics model serving + logging
IOR parallel scaling 8 → 32 clients	detect bottlenecks in scale-out
mdtest metadata	small-file namespace stress
S3 1 MiB PUT/GET, 256 KiB, 8 KiB, multi-part	assess striping & object micro-sharding

Test	Why
NFS-RDMA latency sweep (Linux nfs-latency)	verify low-jitter RDMA path
Fault injection (power-pull NVMe, kill-node)	ensure HA & rebuild budgets

4.3 Deliverables

- JSON/CSV raw metrics committed to repo.
- Grafana dashboard exports.
- Formal PoC report template including traffic captures, system counters, firmware versions.

Exit gate: PoC vendor scores finalized.

5 | Operational Fit & Risk Assessment (2 weeks)

Domain	Checks
Security	SBOM provided, rpkgs signed, vulnerability disclosure window ≤ 30 days
Supportability	24 × 7 L3, onsite parts depot ≤ 4 h; upgrade in-place, rolling
Supply chain	country of origin, second-source NAND, roadmap for 3 yrs
Compliance	ISO 27001, SOC 2 Type II, FedRAMP Moderate (reference CISO checklist)
Integration	Ansible roles installed in dev cluster; monitoring exporters feed Prometheus/Alertmanager

- Run Table-Top Failure Mode & Effects Analysis (FMEA) with vendor TAM present; assign severity/occurrence/detectability scores.

6 | Commercial & Contractual Negotiation (2 weeks)

- TCO model: CAPEX list price -> discounted, maintenance (Yr 1-5), power/cooling at $/kWh, rack U cost.
- Price/performance KPI: $ / 4 KiB IOPS, $ / GB/s, $ / PB effective.
- Legal: indemnity, data-sovereignty clauses, ransomware warranty if offered.
- SLA: < 2 h response, 30 min critical update, firmware upgrade cadence 2× yr.

Deliverable

Generate side-by-side cost sheet and risk register; pass to sourcing review board.

7 | Decision & Handoff (1 week)

1. Compile Evaluation Scorecard – combine phases 3–6 weighted totals → final percentage.
2. Conduct Decision Meeting; minutes archived.
3. Issue Intent to Award.
4. Publish Run-Book for Production Deploy: rack diagram, cabling, IP plan, Ansible playbooks

8 | Post-Deployment Validation (30 days after go-live)

- Compare production telemetry vs PoC baselines (± 10 % tolerance).
- Perform Day-30 Support Review; document lessons learned into knowledge base to close the loop for future refresh cycles.

Appendix A, Vendor Assessment Checklist

Category	Example Evidence	Weight
Technical capability	PoC KPIs met, SPC-1/SPECsfs scores	25 %
Road-map & innovation	PCIe 5.0, CXL readiness, DPUs	10 %
Support & Services	TAM ratio, onsite spares, auto-case-open	15 %
Security posture	SBOM, FIPS, pen-test reports	10 %
Financial health	public filings, Dunn & Bradstreet	10 %
Sustainable operations	Scope 1/2 emissions data, e-waste program	10 %
Commercial terms	TCO, discount, licensing model	20 %