Cithorum 1 PB Commercial Pod — Hardware Specification

Reference rig for the proof environment; carries forward unchanged into the 1 MW build.

Cithorum / Strata India Pvt Ltd · Final spec · 5 May 2026

1. Executive summary

This document specifies the bill of materials for the Cithorum 1 PB commercial pod — the production reference rig for the proof environment. The pod is five identical all-NVMe storage nodes plus one control / GPU node in a single 42U colo cabinet, networked over refurbished 100 GbE Mellanox kit with passive copper DAC cabling. Raw capacity is 1.23 PB; usable capacity after erasure coding (k=4, m=1) is approximately 983 TB. Typical draw is 4–5 kW with peaks around 6 kW. All-in capex sits in the $60–80K range. Every architecture decision in this spec — CPU family, NIC and switch vendor, software stack, erasure-coding scheme — carries forward unchanged into the 1 MW build.

2. Pod composition

Single 42U colo cabinet, 7–9U occupied:

- 5 × storage node — Supermicro AS-1115HS-TNR (1U each)
- 1 × control / GPU node — same chassis (1U)
- 1 × ToR switch — Mellanox SN2700, 32-port 100 GbE (1U)
- 1 × OOB switch — 1 GbE managed, IPMI / BMC (1U)
- 2 × vertical 32A PDUs — A + B feed (colo-provided)

Six servers + ToR + OOB = 7–9U occupied, leaving 30+ U headroom in the cabinet for future expansion.

3. Storage nodes (× 5)

Chassis	Supermicro AS-1115HS-TNR — 1U, 8-bay U.2 NVMe front access
CPU	AMD EPYC 7313P — 16C / 3.0 GHz / 155 W · Milan generation · single socket · “P” SKU
RAM	256 GB DDR4-3200 ECC (4 × 64 GB RDIMMs); upgrade path to 512 GB once paying users justify
Drives	8 × Samsung PM9A3 30.72 TB U.2 NVMe · PCIe Gen 4 · TLC NAND · 1 DWPD endurance
NIC	Mellanox ConnectX-5 dual-port 100 GbE — refurbished, vendor warranty
PSU	Dual hot-plug, A + B feed
Per-node raw capacity	~245.76 TB

Per-node network	200 Gbps total (2 × 100 GbE)
Per-node power	~700 W typical draw

4. Control / GPU node (× 1)

Chassis	Supermicro AS-1115HS-TNR — identical to storage nodes (parts commonality)
CPU	AMD EPYC 7313P — identical to storage nodes
NIC	Mellanox ConnectX-5 dual-port 100 GbE — identical to storage nodes
RAM	256 GB DDR4-3200 ECC; upgrade to 512 GB once GPU workload justifies
GPU	NVIDIA L4 — 24 GB, 72 W · inference workloads · upgrade path to L40S only when training revenue lands
Drives	2 × 1.92 TB NVMe — boot + local cache (no bulk storage on this node)
Hosts	RustFS coordinator · customer dashboards · billing export · audit logs · AI inference · security-ops VM

5. Networking

ToR switch	Mellanox SN2700 — 32-port 100 GbE · refurbished (~$5–8K used vs ~$15–20K new SN3700C)
OOB switch	1 GbE managed (Netgear / TP-Link enterprise, ~$200) for IPMI / BMC access
Cabling — fabric	14 × 100 GbE DAC, passive copper, in-cabinet runs <3 m
Cabling — OOB	7 × Cat6A

6. Rack, power, cooling

Cabinet	1 × 42U colo cabinet (provided in lease)
PDUs	2 × vertical 32A · A + B feed (colo-provided)
Space used	7–9U occupied (6 × 1U servers + 1U ToR + 1U OOB)
Power — typical	4–5 kW pod-level draw
Power — peak	~6 kW
Cooling	Standard colo CRAC; no special envelope. Milan-gen + Gen 4 NVMe selected partly for thermal headroom in Indian colo summers.

7. Pod totals

Metric	Value	Metric	Value
Raw capacity	1.23 PB (5 × 245.76 TB)	Usable capacity	~983 TB after EC k=4,m=1
Erasure coding	k=4, m=1 · 25% overhead · survives loss of any one node	East-west fabric	~1 Tbps (5 × 200 Gbps)
Typical draw	4–5 kW	Peak draw	~6 kW
Capex (all-in)	~$60–80K	Rack footprint	7–9U of 42U

8. Software stack

The software stack is intentionally narrow. Three components matter for the BOM context; everything else is operations tooling layered on top.

Jam codec	Cithorum / Strata IP. Compression and erasure-coded chunk layout (k=4, m=1 at pod scale; k=8, m=2 at 1 MW scale). Runs on every node, in-line on the write path.
RustFS S3 layer	S3-compatible object interface. Coordinator runs on the control node; each storage node runs a RustFS daemon that owns its local NVMe drives. No external metadata service required.
Operating system	AlmaLinux 9 (preferred) or Ubuntu 22.04 LTS. Identical image on every node — storage and control nodes differ only in workload, not in OS build.

Customer dashboards, billing export, audit logs, AI inference and the security-ops VM all run on the control node. None of those services holds bulk customer data; the data plane lives entirely on the five storage nodes.

9. Design rationale

Milan EPYC + Gen 4 NVMe over Genoa + Gen 5

Milan-generation EPYC paired with Gen 4 NVMe runs roughly 20% lower in thermal envelope than Genoa + Gen 5. That headroom matters in Indian colo summers, where ambient and CRAC margins are tighter than in temperate-climate facilities. The Gen 5 cost premium at this drive density is not justified for a compression-bound workload — the network ceiling binds long before drive bandwidth does.

All-NVMe over HDD bulk + NVMe hot tier

Going all-NVMe collapses the tiering question. There is no HDD/NVMe split, no “demo path vs bulk path,” no decision tree at write time. Every node is identical and every node is a hot node. The $/TB premium is absorbed by refurbished networking and DAC cabling. Operational simplicity is itself a cost saving.

Five nodes specifically

Five nodes maps cleanly to EC k=4, m=1: exactly one chunk per node, 25% storage overhead, survives loss of any one node. Three storage nodes would have forced k=2, m=1 (50% overhead) — noticeably less efficient. Five is the minimum useful node count for this scheme.

Refurbished Mellanox networking

The Mellanox SN2700 ToR (~$5–8K used) and ConnectX-5 NICs (refurbished, vendor warranty) deliver the same line-rate 100 GbE performance as a new SN3700C build (~$15–20K) at a meaningful capex saving. The warranty position is unchanged.

DAC over AOC cabling

All node-to-ToR runs are under three metres inside the cabinet, so passive copper DAC is electrically sufficient and is roughly four to five times cheaper than active optical cabling. AOC is reserved for inter-cabinet runs at 1 MW scale.

10. 1 MW scale-up notes

The proof pod's architecture choices carry forward unchanged into the 1 MW reference build. The seed crystal defines the crystal.

What stays the same

- Same CPU family (AMD EPYC) — Milan today, Genoa or Turin at 1 MW only if thermal envelope and Indian colo conditions allow it.
- Same NIC and switch vendor (Mellanox / NVIDIA Networking).
- Same software stack — Jam codec + RustFS S3 layer + AlmaLinux/Ubuntu.
- Same erasure-coding scheme, parameterised — moves from k=4, m=1 at pod scale to k=8, m=2 at 1 MW (lower overhead, higher fault tolerance, available because the node count supports it).

What is added at 1 MW

- Tiered storage: hot NVMe (~5%), warm SSD (~20%), bulk HDD (~75%).
- Leaf-spine network topology in place of single-ToR.
- Formal UPS engineering — VRLA at the conservative 3.6× case, sized down further for backup-tier workloads.
- Captive solar PPA — Suraj's 2,000 MW network at ~₹2.35/unit against the ~₹8/unit grid baseline.
- Formal Tier III audit and certification.

Capex at 1 MW: roughly $8–10M rack-level under the conservative case. The 1 PB pod is the seed crystal that defines the 1 MW build — not a different system, just a denser one.