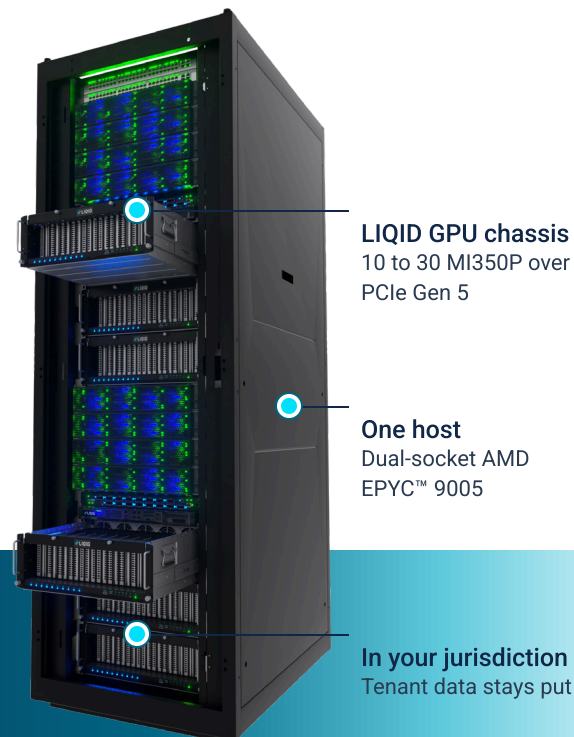


Sell tokens, not idle GPUs.

Sovereign LLM-as-a-service on AMD Instinct™ MI350P. A three-tier model catalog with per-tenant endpoints on one Liquid UltraStack server, in your facility and jurisdiction.



10 GPUs

MI350P on one host, room for 30

1.44 TB

HBM3E on one server

3 tiers

fast, balanced and premium

0 hops

network hops between GPUs

In your jurisdiction
Tenant data stays put

The Capacity Dilemma

Bet on a GPU cluster before tenants sign, or miss the demand for private endpoints.



Demand Pressure

Customers want private endpoints

In-country LLM endpoints their data never leaves.

Demand shifts by tier

Fast, 70B and reasoning traffic rise and fall at different times.



Capital Pressure

A cluster is a big bet

Large GPU clusters come before paying tenants do.

Peak sizing sits idle

Buying for peak leaves capacity unused most of the day.



Sovereignty Pressure

Data stays in-country

Tenant data has to stay in your facility and jurisdiction.

Every tenant is isolated

Each one gets its own endpoint and API keys.

Serve It. Meter It. Grow It.

One 10-GPU Liquid UltraStack server serves a three-tier catalog: a fast small model, a balanced 70B and a premium reasoning model. Each tenant gets its own endpoint, and GPUs shift between tiers in software.

- » Sell tokens, not idle GPUs
- » Low starting risk: one server, not a cluster
- » Grow from 10 to 20 to 30 GPUs on the same host

Per-tenant endpoints and API keys · three-tier catalog

4
Fast · 8B

3
Balanced · 70B

2
Premium · 120B

1
Flex

LIQID GPU chassis · 10 AMD Instinct™ MI350P · 1.44 TB HBM3E

PCIe Gen 5 connection · chassis to host

One dual-socket AMD EPYC™ 9005 host · one OS image

Fast Tier · 4 GPUs

Llama 3.1 8B replicas serve the low-cost, high-volume tier.

18,526

tok/s, 7B-class, 4 GPUs

44,383

tok/s, 7B-class, 10 GPUs

Balanced Tier · 3 GPUs

Llama 3.3 70B, one full model per GPU, for general-purpose endpoints.

1,256

tok/s, 72B-class, 8 GPUs

1 GPU

per 70B model at FP8

Premium + Flex · 3 GPUs

gpt-oss-120b reasoning tier on 2 GPUs; 1 flex GPU for the busiest tier or failover.

144 GB

HBM3E per MI350P

1 card

holds gpt-oss-120b

Measured results: Liquid Phase-3 testing on 1 to 10 MI350P GPUs on one host. Test conditions and notes on page 2.

Built for

Neoclouds · Cloud service providers · Sovereign cloud · Managed service providers

Proven on

AMD Instinct™ MI350P · AMD EPYC™ 9005 · ROCm™ · 200+ Liquid deployments

Size Your Catalog

sales@liquid.com · +1 303.500.1551

Proof: Measured on MI350P

No cluster bet: one server, three tiers, and capacity that grows with paying tenants.

Result	What was measured	Why it matters here
18,526 tok/s on 4 GPUs, 7B-class	FP4, one replica per GPU, fixed concurrency of 128	High-volume fast tier on a fraction of the server
3.98× scaling 1 to 4 GPUs, 7B-class	Near-linear from 1 to 4 GPUs; 44,383 tok/s at 10 GPUs	Every GPU you add is capacity you can sell
552 tok/s on 4 GPUs, 72B dense	FP8 under interactive limits: P99 TTFT ≤ 2 s, ≥ 30 tok/s per user; 1,256 on 8 GPUs	Balanced tier at interactive speed for every tenant

What it means	Capacity matches paying demand	Grow by the GPU, not by the cluster
---------------	--------------------------------	-------------------------------------

Two tests, different conditions: the 7B-class run (FP4) used fixed concurrency of 128 with no latency gate; the 72B run (FP8) reports throughput at the latency-gated knee. Don't compare them directly. Source: Liquid Phase-3 testing. gpt-oss-120b, Llama 3.3 70B and Llama 3.1 8B performance is projected from same-class results and has not been measured directly.

Grow on the Same Server: 10, 20 and 30 GPUs

Add GPUs and a new job, not a new cluster.

10 GPUs
1.44 TB HBM3E

START HERE

Sovereign three-tier LLM catalog
Fast, balanced and premium endpoints with per-tenant API keys.

MODEL LAYOUT
4 × Llama 3.1 8B fast tier
3 × Llama 3.3 70B balanced tier
2 × gpt-oss-120b premium tier
1 × flex or failover

WHO BUYS, WHY NOW
Neocloud / CSP product lead. First tenants outgrow a 4-GPU starter.

20 GPUs
2.88 TB HBM3E

Dedicated tenant endpoints
Shared tiers plus reserved GPUs for enterprise tenants that want dedicated capacity.

MODEL LAYOUT
10 × shared catalog tiers (8B, 70B, gpt-oss-120b)
8 × dedicated tenant endpoints
2 × flex and failover

WHO BUYS, WHY NOW
VP Product / GM, AI cloud. Enterprise tenants asking for dedicated, in-country capacity.

30 GPUs
4.32 TB HBM3E

Sovereign frontier tier
A frontier-class open model alongside the full catalog, in your jurisdiction.

MODEL LAYOUT
8 × Mistral Large 3 (675B, FP8)
20 × catalog tiers and dedicated endpoints
2 × flex and failover

WHO BUYS, WHY NOW
CTO / sovereign cloud program. Customers want frontier models without a hyperscaler.

Starting smaller? The 4-GPU starter catalog runs 1 GPU per tier plus 1 flex. MoE models such as Llama 4 Maverick can span 4 GPUs (576 GB HBM3E); not yet measured.

What UltraStack Adds

- » **One model per GPU:** 144 GB HBM3E holds a full 70B model at FP8, or gpt-oss-120b.
- » **No network between GPUs:** peer-to-peer over PCIe Gen 5 to one EPYC host.
- » **Air-cooled:** about 22 kW for the full 30-GPU system.

Start Fixed, Add Flexibility

- » **One host:** one OS image, one license footprint, one thing to patch.
- » **Room to grow:** 10 to 20 to 30 GPUs on the same host.

What Is Proven Today

- » P2P bandwidth saturates the link; RCCL collectives pass, 1 to 10 GPUs.
- » ROCm integration complete, developed jointly with AMD.
- » 200+ production deployments, 85+ patents.

The Efficiency Case Targets, not measurements*	Up to 75% fewer servers	Up to 2.1× more tokens per dollar	Up to 1.8× better tokens per watt	Up to 50% lower power
--	-----------------------------------	---	---	---------------------------------

Two questions settle the size of the opportunity: **how many tenants want private endpoints today, and which models do they want?**

Bring Us Three Numbers

- 1 **Tenants:** committed and in the pipeline.
- 2 **Mix:** share of fast, 70B and reasoning traffic.
- 3 **Peak:** concurrent users and tokens/sec at peak.

Request a Sizing Session

We size it in one working session.
sales@liquid.com · liquid.com

*Efficiency targets compare against multi-server deployments of the same GPUs; they are Liquid internal projections using AMD published specifications, and results vary by configuration and workload. Performance figures for the AMD Instinct MI350P are based on AMD published specifications and Liquid Phase-3 testing from 1 to 10 GPUs, and are subject to change. Performance at 20 and 30 GPUs, and for 400B+ models, is not yet published; 20- and 30-GPU layouts and frontier-model fit are sizing arithmetic on published model sizes, not benchmark results. Models shown are US and EU open-weight models.