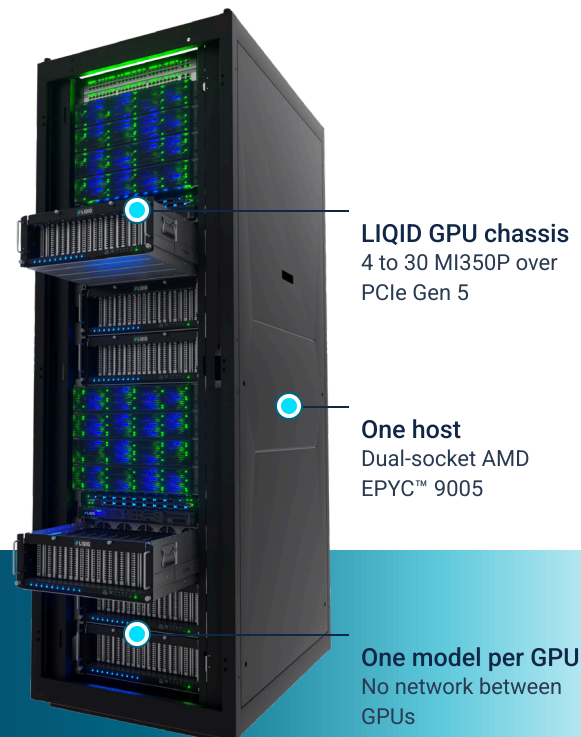


Every model. One server.

Liquid UltraStack with AMD Instinct™ MI350P. Run every model your teams need in parallel, one model per GPU, on a single AMD EPYC™ server. Available now from 4 to 30 GPUs.



30 GPUs

MI350P on one host

4.32 TB

HBM3E on one server

30 models

in parallel, one per GPU

0 hops

network hops between GPUs

One model per GPU
No network between GPUs

The Growth Dilemma

Buy another server for every new job, or run every model your teams need on one.



Model Pressure

Every team wants its own models

Each new job adds models to route, extract, draft and check.

Agents multiply the work

Agents use 10–100× more tokens per task.



Server Pressure

Standard servers fill fast

Each new job means another server to buy, rack and run.

Every node adds overhead

Another OS, more NICs, ports and licenses.



Data Pressure

Data can't leave

Regulated and confidential data can't go to a hosted API.

Pinned models are easier to defend

Self-hosted models don't change under you.

One Model per GPU. One Host.

Liquid GPU Pooling attaches 4 to 30 MI350P GPUs to one AMD EPYC host over PCIe Gen 5. Each GPU runs its own model: 20 GPUs, 20 models at once.

- » Available now from 4 to 30 GPUs
- » Fixed, tested configurations
- » Grow by adding GPUs, not servers

Liquid UltraStack with AMD MI350P · one model per GPU

4

One job

10

Team

20

Department

30

Enterprise

LIQID GPU chassis · up to 30 AMD Instinct™ MI350P · up to 4.32 TB HBM3E

PCIe Gen 5 connection · chassis to host

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

Many Models at Once

Each model runs on its own GPU, so capacity grows in a straight line.

9.5×

scaling, 1 to 10 GPUs

~89K

tok/s at 20 GPUs

Every Model on One Card

144 GB of HBM3E holds a 70B model at FP8 or gpt-oss-120b.

144 GB

HBM3E per MI350P

~120B

class model per GPU

Fits Your Facility

Standard PCIe cards, air-cooled, on a server you know.

600 W

per GPU, air-cooled

~22 kW

at 30 GPUs

Measured results: Liquid Phase-3 testing on 1 to 20 MI350P GPUs on one host. 30-GPU figures are projected. Notes on page 3.

Built for

Multi-model serving · Agentic AI · RAG and document Q&A · Private AI · Neocloud catalogs

Proven on

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

Size Your System

sales@liquid.com · +1 303.500.1551

Configurations: 4 to 30 GPUs on One Server

Every configuration is **available now**, and every one runs one model per GPU.

Configuration	4 GPUs	10 GPUs	20 GPUs	30 GPUs
Tier	One job	Team	Department	Enterprise
Models in parallel (one per GPU)	4	10	20	30
Aggregate HBM3E	576 GB	1.44 TB	2.88 TB	4.32 TB
FP8 AI performance	9.2 PF	23 PF	46 PF	69 PF
7B-class throughput	18,526 measured	44,383 measured	~89,000 measured	~133,000 projected
GPU power (600 W per GPU)	2.4 kW	6 kW	12 kW	18 kW
Liquid 10-slot GPU chassis	1	1	2	3
Network hops between GPUs	0	0	0	0

7B-class: FP4, one model per GPU, concurrency 128, tokens/s, not latency-gated. Any count from 4 to 30 GPUs is available; Liquid sales leads with 10, 20 and 30, and partners start at 4.

What You Can Run as You Grow

Add GPUs and the next job's models, not a new server.

10 GPUs

1.44 TB HBM3E · Team

Every job one team runs

Ten models side by side, one per GPU.

EXAMPLE MODEL LAYOUT

4 × Llama 3.1 8B triage · 4 × Llama 3.3 70B drafting · 2 × gpt-oss-120b QA

EXAMPLE INITIATIVES

Financial crimes investigation · clinical documentation · practice-group review · analyst Q&A

20 GPUs

2.88 TB HBM3E · Department

A department's full set of models

Twenty models running in parallel.

EXAMPLE MODEL LAYOUT

Five agents × 4 GPUs: prior auth, appeals, eligibility, claims triage, denials

EXAMPLE INITIATIVES

Revenue cycle suite · defensible e-discovery · FOIA and redaction · dedicated tenant models

30 GPUs

4.32 TB HBM3E · Enterprise

An enterprise-wide private model hub

Every department's models on one host.

EXAMPLE MODEL LAYOUT

10 × team jobs · 14 × department suite · 6 × shared reasoning and Q&A

EXAMPLE INITIATIVES

Bank-wide model hub · health-system hub · firm-wide hub · sovereign model catalog

Example layouts use US and EU open-weight models that fit on one GPU: Llama 3.1 8B, Mistral Small 3.2, Gemma 3 27B, Llama 3.3 70B and gpt-oss-120b. Layouts are sizing arithmetic, not benchmark results.

What UltraStack Adds

- » **One model per GPU:** no model shares a card, so performance stays predictable.
- » **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, Grow on One Host

- » **One host:** one OS image, one license footprint, one thing to patch.
- » **Room to grow:** 4 to 10 to 20 to 30 GPUs on the same server.

What Is Proven Today

- » Near-linear scaling measured from 1 to 20 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

Technical Specifications

Standard parts: **your server, Liquid fabric and chassis, AMD GPUs.**

Specification	Liquid UltraStack with AMD MI350P (4 to 30 GPUs)
Description	AI inference server running one model per GPU, 4 to 30 GPUs on one host
Host server	Dual-socket AMD EPYC™ 9005 Series (Dell or Supermicro)
GPUs	4 to 30 × AMD Instinct™ MI350P PCIe (144 GB HBM3E, 600 W, air-cooled per GPU)
GPU memory	576 GB to 4.32 TB HBM3E · ~4 TB/s per GPU
AI performance	9.2 to 69 PFLOPS FP8 · up to 138 PFLOPS MXFP4/MXFP6 (peak, AMD published)
GPU pooling fabric	Liquid PCIe Gen 5 fabric with Liquid host bus adapter; peer-to-peer GPU communication
Expansion chassis	Liquid EX-Series 10-slot GPU chassis (1, 2 or 3; one per 10 GPUs)
Management	1 × Liquid Director 1U management appliance
Software	AMD ROCm™ with vLLM and SGLang through AMD Inference Microservices; runs your existing stack on the host
Storage and networking	Liquid NVMe SSD; high-speed Ethernet or InfiniBand NICs (recommended, configurable)
GPU power	2.4 to 18 kW (600 W per GPU, configurable to 450 W)
Total system power	Scales with GPU count; ~22 kW at 30 GPUs
Cooling	Air-cooled; no liquid cooling or facility redesign

Storage and networking devices are recommended and can be customized or omitted. Power reflects typical AI inference load; actual consumption varies by configuration and workload.

Per-GPU Building Block

Every GPU Adds	Value
One model, up to ~120B-class	e.g. Llama 3.3 70B at FP8 or gpt-oss-120b at MXFP4
HBM3E per GPU	144 GB
FP8 per GPU	~2.3 PFLOPS
Power per GPU	600 W (configurable to 450 W)

Two questions settle the size:
which jobs do you need to run,
and how many users or agents
run them at once?

Bring Us Three Numbers

- 1 Jobs:** the workflows you want on one server.
- 2 Models:** the model for each part of each job.
- 3 Concurrency:** users and agents 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 20 GPUs, and are subject to change. 30-GPU figures are projected from measured 20-GPU scaling. Every configuration runs one model per GPU. Models shown are US and EU open-weight models. 7B-class: 4,653 tok/s on 1 GPU, 18,526 on 4, 44,383 on 10, ~89,000 on 20 (measured); ~133,000 on 30 (projected). AMD specifications from the AMD Instinct MI350P product brochure.