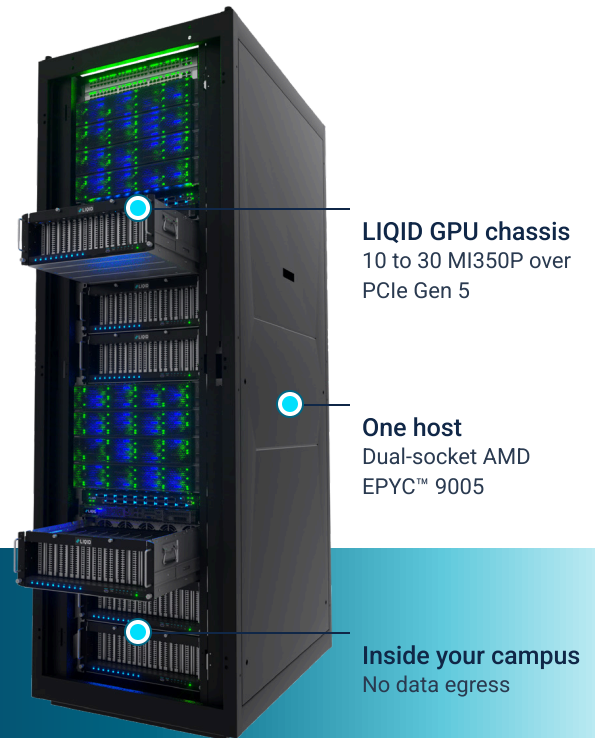




One AI service. Data stays on campus.

# Every lab. One AI service.

**Campus-wide private AI on AMD Instinct™ MI350P.** Every lab gets its own model on one Liquid UltraStack server, with research data kept inside your data center.



**10 GPUs**

MI350P on one host, room for 30

**1.44 TB**

HBM3E on one server

**6 GPUs**

of lab endpoints, one model each

**0 hops**

network hops between GPUs

## The Pilot-to-Campus Dilemma

Keep private AI with one research center, or scale it to every lab without building a cluster.



### Demand Pressure

#### Every lab wants AI

Departments ask for private model access at the same time.

#### Unsanctioned API use

Grant money and protected data drift to public APIs.



### Infrastructure Pressure

#### GPUs locked in silos

Lab-owned servers sit idle while other labs wait.

#### Clusters add overhead

Multi-node builds need GPU networking, switches and staff.



### Compliance Pressure

#### Protected data can't leave

IRB protocols and data use agreements bar public AI APIs.

#### Findings must be traceable

Every coded theme has to link back to its source.

## Share It. Assign It. Reassign It.

One 10-GPU Liquid UltraStack server runs a campus research AI service. Each lab or project gets a full GPU with its own model, and GPUs move to new projects each term or grant cycle.

- » Research data stays on campus
- » One model per GPU, no noisy neighbors
- » Fixed capacity, no per-token fees on grants

Open-weight models · one model per GPU

**6**

Lab endpoints

**2**

Analysis

**2**

Batch coding

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

### Lab Endpoints · 6 GPUs

gpt-oss-120b, one full model per GPU, assigned to a lab or project.

**~65 GB**

fits one card

**1,256**

tok/s on 8 GPUs, 72B

### Analysis · 2 GPUs

Llama 3.3 70B codes themes and summarizes, citing the source.

**254**

tok/s on 2 GPUs

**~8**

live researchers\*

### Batch Coding · 2 GPUs

Llama 3.1 8B labels survey and free-text responses at volume.

**9,193**

tok/s on 2 GPUs

**9.5×**

scaling, 1 to 10 GPUs

Measured: Liquid Phase-3 testing on 1 to 20 MI350P GPUs on one host. Named models are projected. \*Illustrative; notes on page 2.

#### Built for

Research computing centers · Human-subjects research · Campus-wide AI services

#### Proven on

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

**Size Your Campus**

sales@liquid.com · +1 303.500.1551

## Proof: Measured on MI350P

No cluster, no idle silos: **one host, ten GPUs, a model per lab, and room to grow to 30.**

| Result                                        | What was measured                                                                             | Why it matters here                             |
|-----------------------------------------------|-----------------------------------------------------------------------------------------------|-------------------------------------------------|
| <b>~65 GB</b><br>gpt-oss-120b at MXFP4        | Published model size: the full model fits on one 144 GB MI350P with room for cache            | Six lab endpoints, no model split across cards  |
| <b>1,256 tok/s</b><br>on 8 GPUs               | 72B-class model, FP8, under interactive limits: P99 TTFT $\leq 2$ s, $\geq 30$ tok/s per user | Reference for interactive speed at campus scale |
| <b>9.5× scaling</b><br>1 to 10 GPUs, 7B-class | 4,653 tok/s on 1 GPU to 44,383 tok/s on 10, one replica per GPU                               | Every added GPU adds capacity for the next lab  |
| What it means                                 |                                                                                               | One host, no GPU network to run                 |

Model size is a published spec. 72B-class: FP8, one replica per GPU (254 tok/s on 2 GPUs, 552 on 4, 1,256 on 8). 7B-class: FP4, one replica per GPU, concurrency 128 (4,653 tok/s on 1 GPU, 9,193 on 2, 44,383 on 10). Different test conditions; don't compare them.  $\sim 8 = 254 \div 30$  tok/s per user, illustrative. Source: Liquid Phase-3 testing. Named-model 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.

|                                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>10 GPUs</b><br>1.44 TB HBM3E<br><b>Private AI service for many labs</b><br>One GPU per lab or project, reassigned each term or grant cycle.<br><b>MODEL LAYOUT</b><br>6 × gpt-oss-120b lab endpoints<br>2 × Llama 3.3 70B analysis<br>2 × Llama 3.1 8B batch coding<br><b>BUILT FOR</b><br>CIO / Director of Research Computing.<br>Labs asking for AI; unsanctioned API use on grants. | <b>20 GPUs</b><br>2.88 TB HBM3E<br><b>Research administration suite</b><br>Grant compliance review, export-control screening and proposal drafting.<br><b>MODEL LAYOUT</b><br>6 × 70B proposal and compliance review<br>4 × 8B export-control screening<br>10 × lab endpoints and analysis<br><b>WHO BUYS, WHY NOW</b><br>VP Research / Sponsored Programs.<br>Proposal volume; review backlog. | <b>30 GPUs</b><br>4.32 TB HBM3E<br><b>University-wide private model hub</b><br>Research, administration and teaching on one host.<br><b>MODEL LAYOUT</b><br>14 × lab endpoints and analysis<br>10 × research administration<br>6 × gpt-oss-120b courses and coursework<br><b>WHO BUYS, WHY NOW</b><br>CIO / Provost. Consolidating departmental AI; teaching demand. |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

One host with 30 GPUs serves research, administration and teaching. Starting with one research center? See the 4-GPU Higher Education partner brief.

### 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, Grow on One Host

- » **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.

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

Two questions settle the size of the opportunity: **how many labs are asking for AI today, and how many hold data they can't send to a public API?**

### Bring Us Three Numbers

- 1 Labs:** labs and active projects asking for AI.
- 2 Data:** projects with restricted data and their volume.
- 3 Users:** researchers working 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 layout runs one model per GPU; 20- and 30-GPU layouts are sizing arithmetic on published model sizes, not benchmark results. Models shown are US and EU open-weight models. Outputs support, and do not replace, researcher review. Institutions remain responsible for IRB, data use agreement and export-control compliance.