

Performance Benchmarks

Liquid UltraStack with AMD Instinct™ MI350P GPUs

Add a GPU. Add a model. Liquid UltraStack with AMD Instinct™ MI350P runs one model per GPU, with throughput that scales linearly to 30 GPUs on one AMD EPYC™ server: **measured to 20, projected to 30.**

The LIQID GPU chassis holds 4 to 30 MI350P cards over PCIe Gen 5, pooled behind one dual-socket AMD EPYC™ 9005 host. Every GPU stays in one node, so there are no network hops between GPUs and nothing waits on another card.

~89_k tok/s, 7B-class, 20 GPUs Measured, Liquid Phase-3 testing	~133_k tok/s, 7B-class, 30 GPUs Projected from measured 20-GPU scaling
95% of linear scaling, 1 to 10 GPUs 9.5x throughput from 1 to 10 GPUs	0 network hops between GPUs One host, one fabric, no cluster to manage

How We Tested

Liquid Phase-3 testing on AMD Instinct MI350P, one host, AMD ROCm™ software.

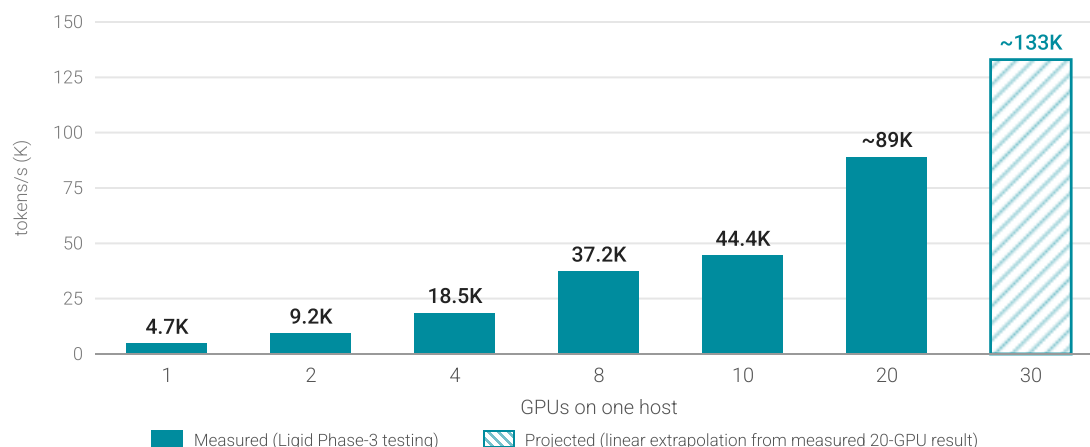
- » **The system:** one host, up to 20 GPUs. MI350P on a dual-socket EPYC host, PCIe Gen 5. One model per GPU; no model was split across cards.
- » **Throughput test:** 7B-class model, FP4, concurrency 128, output tokens/s. Measured 1 to 20 GPUs; 30 GPUs projected.
- » **Interactive test:** 72B-class model, FP8, P99 TTFT ≤ 2 s and ≥ 30 tok/s per user. Measured 2 to 8 GPUs; larger counts projected.

Linear to 30 GPUs

Each GPU runs its own model, so nothing waits on another GPU. Measured linear to 20 GPUs; the same line reaches **~133,000 tok/s at 30**.

- » **9.5× from 1 to 10 GPUs**, 95% of linear.
- » **~89,000 tok/s at 20 GPUs**, measured.
- » **~133,000 tok/s at 30 GPUs**, projected.

7B-class model · output tokens/s, one model per GPU



Performance at 4, 10, 20 and 30 GPUs

The same host at every step: **add GPUs, add models**.

What it delivers	4 GPUs One job	10 GPUs Team	20 GPUs Department	30 GPUs Enterprise
Models in parallel (one per GPU)	4	10	20	30
7B-class output throughput	18,526 ●	44,383 ●	~89,000 ●	~133,000 ○
7B-class tokens per hour	~67M ●	~160M ●	~320M ●	~480M ○
72B-class throughput under latency SLA	552 ●	~1,570 ○	~3,140 ○	~4,710 ○
Interactive users at ≥30 tok/s each	~18	~52	~105	~157
Network hops between GPUs	0	0	0	0

● Measured, Liqid Phase-3. ○ Projected: linear extrapolation from measured results (7B-class from 20 GPUs; 72B-class from 1,256 tok/s on 8 GPUs).
Interactive users = 72B-class throughput ÷ 30 tok/s per user; actual concurrency depends on prompt length, output length and context.

What It Means for Your Workloads

Many Models at Once

Small and mid-size model fleets scale with every GPU.

44,383

tok/s at 10 GPUs

~89K

tok/s at 20 GPUs

Interactive 70B-Class

A 72B-class model held latency targets as GPUs were added.

1,256

tok/s at 8 GPUs

≤ 2 s

P99 first token

Every Model on One Card

144 GB of HBM3E holds the model and its cache.

70 GB

Llama 3.3 70B
FP8

~65 GB

gpt-oss-120b

Two tests, two conditions: compare within a test, not across.

Models You Can Run

Every model below fits on one MI350P with room for its KV cache.

Open model (US / EU)	Precision	Fits one GPU	On UltraStack
OpenAI gpt-oss-120b	MXFP4	~65 GB of 144 GB	One model per GPU; up to 30 on one host
Meta Llama 3.3 70B Instruct	FP8	~70 GB of 144 GB	~13 concurrent 32K sessions per GPU
Google Gemma 3 27B	BF16	~54 GB of 144 GB	Standard-tier chat and summarization
Mistral Small 3.2 24B Instruct	FP16	~48 GB of 144 GB	Classification, extraction, translation
Meta Llama 3.1 8B Instruct	FP8	~8 GB of 144 GB	Routing and triage at high volume

AMD single-GPU (TP1) testing with AMD Inference Microservices (AIM) on ROCm for gpt-oss-120b, Llama 3.3 70B, Gemma 3 27B and Mistral Small 3.2; Llama 3.1 8B performance is projected from 7B-class results. Weights ≈ 1 byte per parameter at FP8, 2 at BF16/FP16, ~0.55 at MXFP4. Concurrent sessions are KV-cache arithmetic (~0.16 MB per token, FP8), not benchmarks.

Why It Scales

- » **One model per GPU.** No model is split across cards, so nothing waits on another GPU.
- » **One node.** No inter-server network or per-node software stack.
- » **Peer-to-peer fabric.** GPU-to-GPU traffic saturates the PCIe Gen 5 link without passing through the CPU. RCCL collectives pass, 1 to 10 GPUs.
- » **Headroom on the host.** Host CPU and memory stay lightly loaded as GPUs are added. Air-cooled: about 22 kW for the full 30-GPU system.

What Next?

The fastest proof is your own workload: **which jobs, which models, and how many users at once?**

Bring us three numbers and we will size it in one working session.

- » **Jobs:** the workflows you need to run.
- » **Models:** the model for each part of each job.
- » **Concurrency:** users and agents at peak, and context length.

Request a sizing session: contact sales@liquid.com or visit www.liquid.com.

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

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

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 tokens/s on 1 GPU, 9,193 on 2, 18,526 on 4, 37,216 on 8, 44,383 on 10, ~89,000 on 20. 72B-class: 254 tokens/s on 2 GPUs, 552 on 4, 1,256 on 8. Tokens per hour = tokens/s × 3,600. AMD Instinct MI350P specifications: 144 GB HBM3E, ~4 TB/s, ~2.3 PFLOPS FP8, 600 W (AMD published).

Liquid Inc.

11400 Westmoor Circle, Suite 225, Westminster, CO 80021

Office: +1 303.500.1551 · sales@liquid.com · www.liquid.com

© 2026 Liquid Inc. All rights reserved. Patent No. 8,688,926. Patents pending. Liquid, Liquid Matrix and UltraStack are trademarks of Liquid Inc. AMD, AMD Instinct, AMD EPYC, ROCm and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other names are for informational purposes only and may be trademarks of their respective owners. Performance information and numbers subject to change.