Bentaus

BENTAUS NEOCLOUD · WHITE PAPER · OCTOBER 2026

Where the AMD Instinct™ MI325X Wins

Production inference on the Bentaus NeoCloud, measuring MI325X against NVIDIA H100: throughput, memory, concurrency, agents, economics and power. Running the Ziani Power Asset Orchestrator.

29,807
output tok/s, Llama 70B FP8 as eight 1-GPU copies, a layout the H100 can't run

2.22×
node throughput vs H100 once memory becomes the limit

4×
agents held within the performance target (Qwen3-30B)

900K
token prompt on a single MI325X, with 4.96M tokens of KV cache

Independent study by Bentaus. All data points measured solely by Bentaus.

SUMMARY

Finding the best customer use cases for accelerators

At Bentaus, we don't evaluate an accelerator by a single number. As an energy-first NeoCloud, we need to know where a GPU belongs in production: which models, which layouts, how many users and agents, at what price, and at what power.

Bentaus put an 8-GPU AMD Instinct MI325X node and an 8-GPU NVIDIA H100 node through the same production-style tests: AMD's maximum-throughput benchmark, user ramps, multi-step agent tasks, long context, model layouts, precision and power caps. Both nodes used the same model files, the same vLLM version, the same prompts and the same ramps.[B1–B4]

On common benchmarks both GPUs performed well, and at light and moderate load the H100 was often faster per user. But common benchmarks don't capture where AI demand is heading. So Bentaus looked for the workloads that matter to our customers: many concurrent users, multi-step agents, long context and large models. Run side by side on those workloads, the MI325X shines, and it's the GPU most underestimated by a standard benchmark. Its 256 GB of HBM3E per GPU becomes usable capacity: layouts the H100 can't run, sessions it can't hold, and throughput it can't reach.[A1] At $2.00 per GPU-hour, the MI325X delivered more output per dollar than the H100 at $2.69 in every matched workload comparison reported in this study.

29,807
tok/s on Llama 70B FP8 as eight 1-GPU copies: 1.67× the H100's best layout

2.22×
node throughput on Qwen3-30B once the H100's memory fills: 59,026 vs 26,590 tok/s

4×
agents held within the performance target on Qwen3-30B BF16, with 1.62× the tasks per hour

5,292
tok/s on Llama 3.1 405B FP8 at max load: 1.53× the H100

4.6×
full 128K-token sessions per node: 65 vs 14

97%
of output on 23% less GPU power, energy per token −20%: Ziani efficient power management

Where the H100 performs better. On Llama 70B interactive serving the H100 was 15–22% faster at every user count we tested. It was also faster on 405B at 64–96 users, on the Qwen3-30B FP8 user ramp, and on Llama 70B agents. It uses less energy per token in most uncapped tests. Section 07 covers each case.

Independent study by Bentaus. All data points measured solely by Bentaus.

HIGHLIGHTS

The results at a glance

> **The MI325X becomes especially strong when HBM capacity, concurrency, long context, model residency and price become first-order constraints.** Bentaus then uses Ziani to optimize the power required for that useful compute.

Test basis: 8 × MI325X, 256 GB HBM3E per GPU, vLLM 0.23.0, ROCm 7.14.1, $2.00 per GPU-hour. Comparison: 8 × H100 80 GB, same vLLM version and models, $2.69 per GPU-hour.

THROUGHPUT
29,807 tok/s
Llama 70B FP8 as eight 1-GPU copies, 1.67× the H100's best layout. The H100 can't run this layout.

AGENT THROUGHPUT
1.62× tasks/hr
2,597 vs 1,600 Qwen3-30B agent tasks per hour within the performance target.

CONCURRENCY
59,026 tok/s
Qwen3-30B at 4,096 users per node, 2.22× the H100's peak, with 42% of KV cache used.

AGENTS
4× the agents
held within the performance target on Qwen3-30B: $0.25 vs $1.35 per agent-hour.

LONG CONTEXT
65 vs 14
full 128K-token sessions resident per node: $0.25 vs $1.54 per session-hour.

LARGE MODELS
5,292 tok/s
on Llama 3.1 405B FP8 at max load: 1.53× the H100, and 2.06× the tokens per dollar.

EXTREME CONTEXT
900K tokens
prompt accepted on one GPU, with 4.96M tokens of KV cache on that GPU.

ENERGY
0.25 J/token
at max load with one copy per GPU, vs 0.27 for the H100's best layout.

ZIANI
97% output
on 23% less GPU power, energy per token −20% (Llama 70B, 700 W cap). On agents: +33% tasks per MW at 600 W.

Independent study by Bentaus. All data points measured solely by Bentaus.

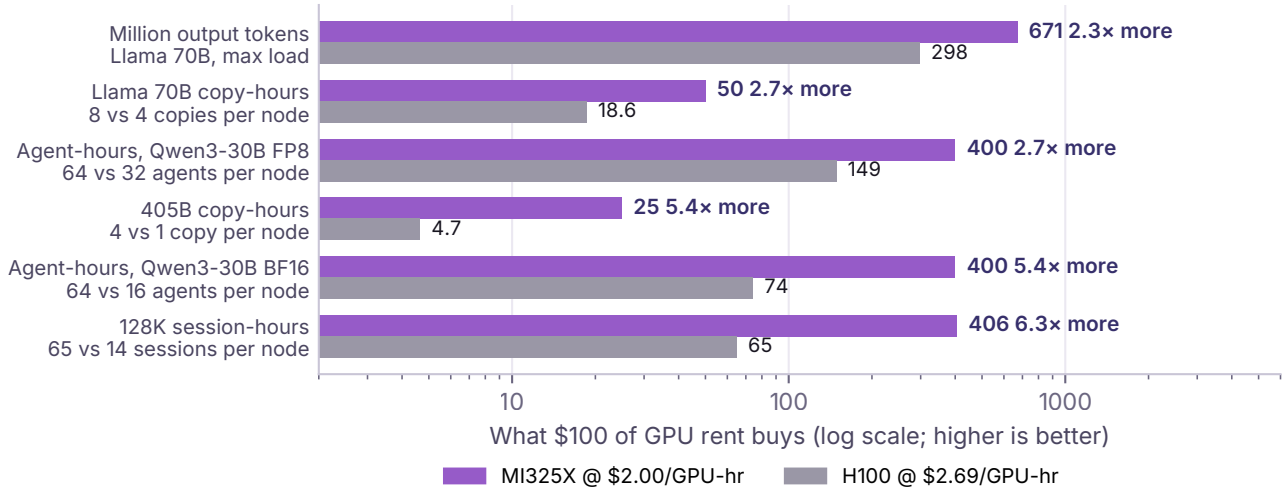

Figure H. What $100 of GPU rent buys on each GPU, computed from measured capacity per node (MI325X node $16.00/hr, H100 node $21.52/hr). On every measure the MI325X delivers 2.3× to 6.3× more per dollar: it costs less per hour and holds more per node. Sources: B2, B4, B5, B6.

Definitions. An **agent-hour** is one agent actively running multi-step tasks for an hour, at the highest agent load where at least 90% of tasks met the performance target. A **session-hour** is one full 128K-token context held resident in GPU memory for an hour, from vLLM's reported KV-cache capacity. A **copy-hour** is one complete model replica serving for an hour. Token figures are output tokens at maximum load.

01 · PEAK THROUGHPUT

The GPU is faster than its reputation

We started with AMD's published vLLM benchmark: Llama 3.1 70B FP8, 128 input and 2,048 output tokens, 3,200 requests at once. We copied AMD's container image, FP8 model and MAD benchmark flags exactly. [A2, A3, A5, A6] AMD publishes this result only for MI300X (13,562 tok/s) and no vLLM inference figure for MI325X.[A2]

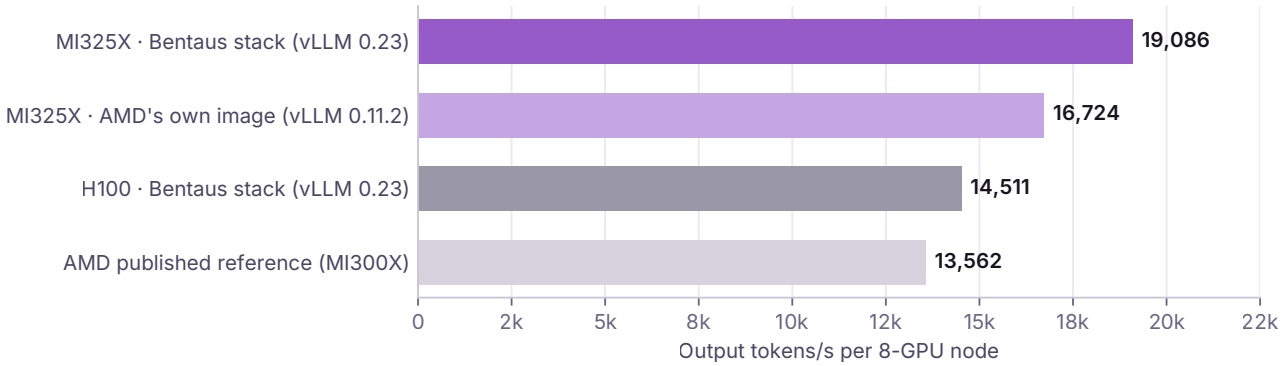

Figure 1. AMD's benchmark, one copy across 8 GPUs, offline maximum load. MI325X runs are averages of three (spread ≤1.1%). The AMD figure was measured on MI300X and is a reference, not a like-for-like result. Sources: B1, B2, A2.

On AMD's own image our MI325X node produced **16,724 tok/s**, 23% more tokens than the MI300X reference. On the stack Bentaus runs in production (vLLM 0.23, ROCm 7.14.1), the same hardware reached **19,086 tok/s**, against 14,511 for the H100 on the same vLLM version.[B1, B2]

Independent study by Bentaus. All data points measured solely by Bentaus.

That benchmark splits one copy of the model across all eight GPUs. The MI325X doesn't need to. With 256 GB per GPU it can run eight independent copies, one per GPU, with no GPU-to-GPU traffic. That layout produced **29,807 tok/s**. The H100 can't run it: 67.7 GB of weights leave no room for user context on an 80 GB GPU.[B2, B4]

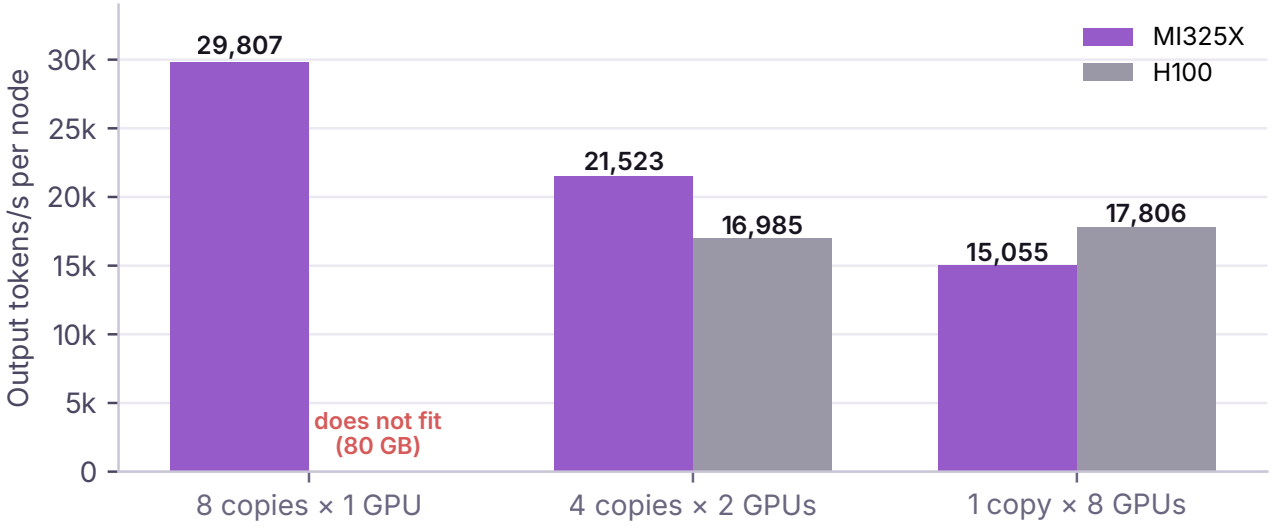

Figure 2. Llama 3.1 70B FP8 at maximum load (128 in / 2,048 out, served) by layout. Best vs best: 29,807 vs 17,806 tok/s, 1.67×, and 2.25× the tokens per dollar. Source: B4, B2.

02 · MEMORY

Memory decides what you can run

Every serving system splits GPU memory between model weights and the KV cache, the working memory that holds each user's context. When weights leave little room, a GPU either serves fewer users or has to be grouped with others. On the H100, that happens early.

Configuration	MI325X (256 GB)	H100 (80 GB)
Llama 3.1 70B FP8 on one GPU	Runs · 1.06M KV tokens per GPU	Does not fit · no room for KV cache
Llama 3.1 405B FP8 on two GPUs	Runs · four copies per node	Does not fit
Llama 3.1 405B FP8 on four GPUs	Runs · 2.1M KV tokens per copy	Does not fit · 320 GB < 410 GB
Llama 3.1 405B FP8 on eight GPUs	Runs · 5.33M KV tokens	Runs · 0.75M KV tokens
Qwen3-30B BF16 at full context on one GPU	Runs · 1.86M KV tokens	Cannot start at its 262K default
Llama 4 Scout FP8, 900K-token prompt, one GPU	Runs · 4.96M KV tokens	Weights alone exceed 80 GB

Independent study by Bentaus. All data points measured solely by Bentaus.

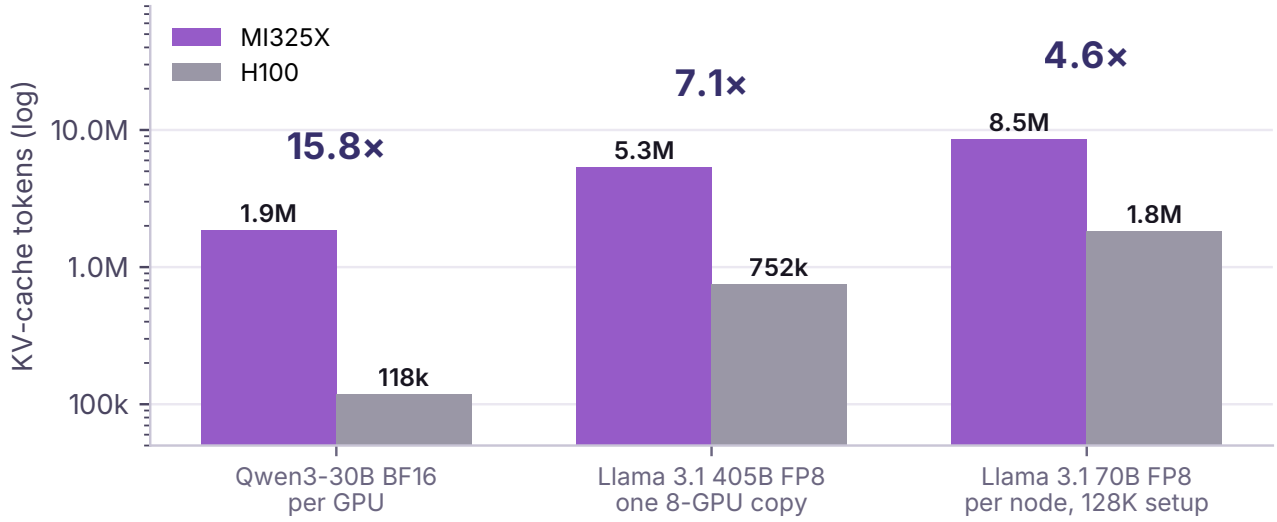

Figure 3. KV-cache capacity reported by vLLM at startup, same model and settings on both GPUs. Sources: B2, B4, B5.

03 · CONCURRENCY

When memory becomes the bottleneck, the curve changes

We ramped Qwen3-30B-A3B (BF16) from 1 to 512 concurrent users per GPU, one independent copy per GPU, at 1,024 input and 512 output tokens. The performance target was time to first token ≤2 s and inter-token latency ≤200 ms, both P99.[B4, O2]

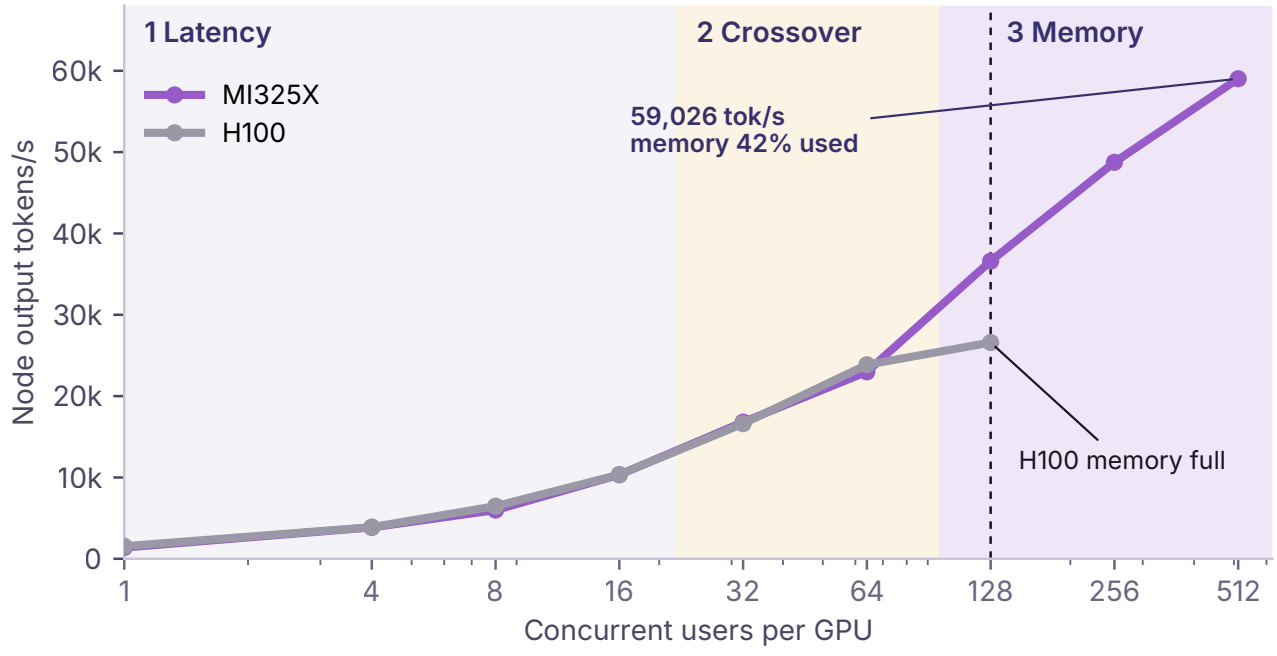

Figure 4. Node output throughput as users grow, Qwen3-30B-A3B BF16, eight independent copies per node. Both GPUs met the performance target up to 64 users per GPU. Source: B4.

Independent study by Bentaus. All data points measured solely by Bentaus.

1 · Latency zone

Up to ~16 users per GPU, the H100 matches or slightly leads.

2 · Crossover zone

32–64 users per GPU. Speed is tied (22,987 vs 23,855 tok/s at 64). Price gives the MI325X 1.3× the tokens per dollar.

3 · Memory zone

128+ users per GPU. The H100's KV cache is full at 128 users per GPU and its throughput stops at 26,590 tok/s. The MI325X keeps climbing to **59,026 tok/s at 512 users per GPU, 2.22× the H100 and ~3× the tokens per dollar**, with 42% of its KV cache used.

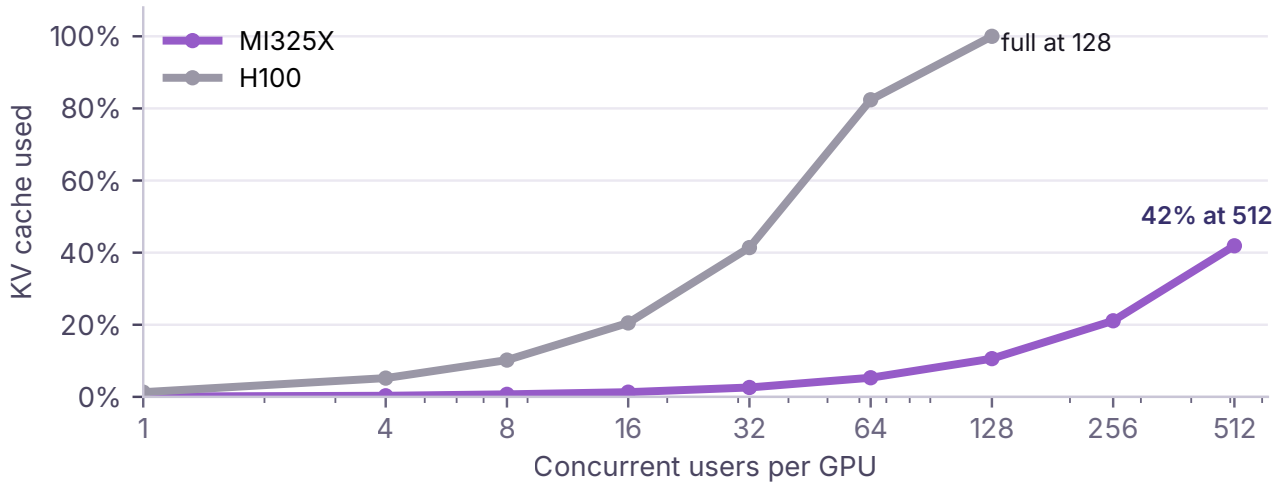

Figure 5. KV cache used as users grow, same test. At 512 users per GPU the MI325X uses about as much of its memory as the H100 does at 32. Source: B4.

Precision matters here. At FP8, the H100 node is faster on this ramp: 48,098 vs 38,162 tok/s within the performance target at 128 users per GPU, and 60,124 vs 49,047 at peak. The MI325X stays 7–10% ahead per dollar. AMD's tuned MoE kernels don't yet run this model on our stack, so MI325X FP8 ran without them. For this model on MI325X, BF16 is currently the faster choice, and it reaches what the H100 reaches at FP8.

04 · AGENTS

From tokens to completed work

Agent customers pay for finished tasks, not tokens. Each task in our test is 15 model calls: a shared 2,000-token system prompt and a 2,000-token task, then 256 generated tokens and a 3,500-token tool result per call. Context grows to about 60,000 tokens per agent. A task meets the performance target only if every call starts within 3 s and streams at ≤200 ms per token. A load level counts only if at least 90% of its tasks meet it. Every run uses fresh tasks, so no GPU benefits from reusing cached work.[B4]

Independent study by Bentaus. All data points measured solely by Bentaus.

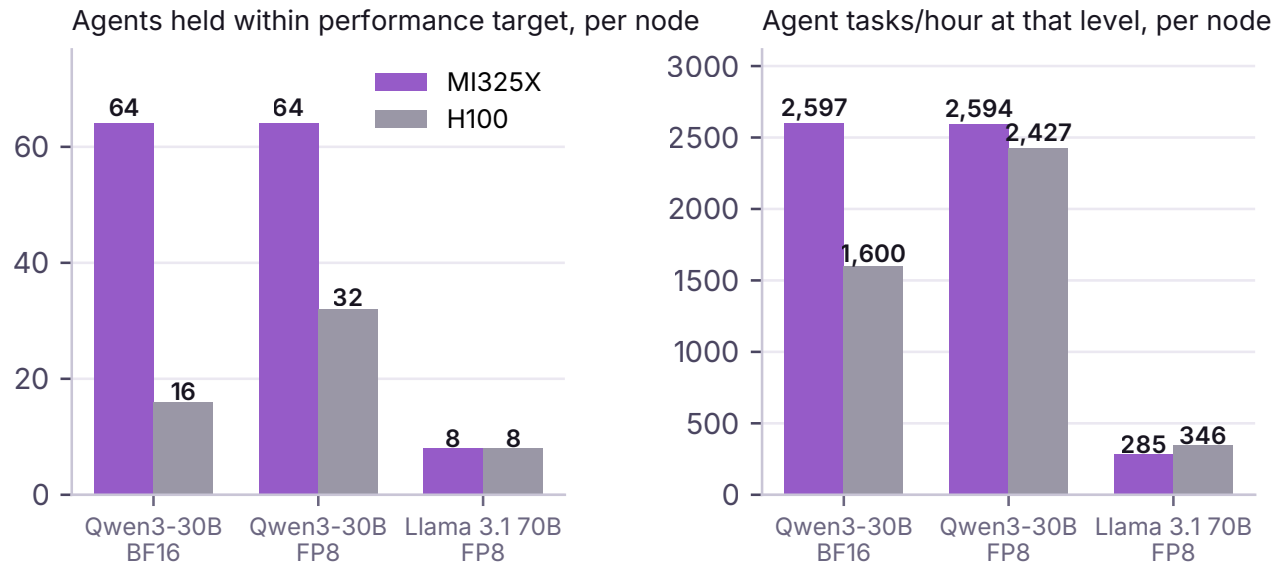

Figure 6. Highest agent load per node where ≥90% of tasks met the performance target, and tasks completed per hour at that load. MI325X used AMD's AITER attention backend. Source: B4, B6.

On Qwen3-30B, the MI325X held **64 agents per node** within the performance target at both precisions. The H100 held 16 at BF16 and 32 at FP8, then ran out of memory. At BF16 the MI325X completed **2,597 tasks per hour against 1,600: 1.62× the throughput and 2.18× the tasks per dollar.** At FP8 it was 2,594 against 2,427: 1.44× per dollar.[B4, B6]

On Llama 3.1 70B, neither GPU kept 90% of tasks within target above one agent per GPU. At that level the H100 completed 346 tasks per hour against 285; the MI325X stays 11% ahead per dollar. Larger agent models need more per-step speed than this test allowed, on both platforms.

> **New metrics for agentic infrastructure.** Bentaus measures agentic capacity as **concurrent agents per dollar**, **completed agent tasks per dollar** and, with Ziani, **completed agent tasks per MW**. Tokens are the input. Finished work is the output.

05 · LONG CONTEXT

When context grows, HBM becomes infrastructure

We sent Llama 3.1 70B FP8 prompts of 30,000, 60,000 and 100,000 tokens and added sessions until memory filled. One H100 can't hold the model with room for context, so it ran as four 2-GPU copies. The MI325X ran as eight 1-GPU copies.[B2, B4]

Independent study by Bentaus. All data points measured solely by Bentaus.

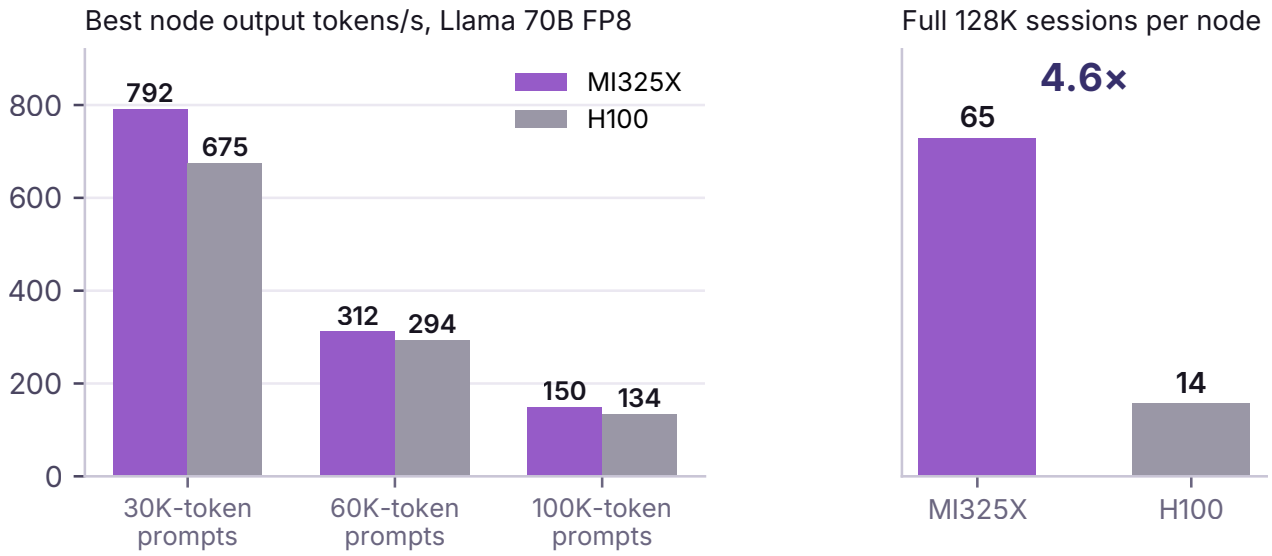

Figure 7. Left: best node throughput at each prompt length. Right: full 128K-token sessions per node, from vLLM's startup report (8.51M vs 1.83M KV tokens). Sources: B2, B4, B5.

Processing long prompts is compute-heavy, so throughput gains are moderate: 1.06–1.17×, or 1.4–1.6× per dollar. Capacity is where the MI325X separates: **65 full 128K-token sessions per node against 14.** At one or two sessions per GPU, the H100 returns the first token sooner.

> **900,000 tokens. One GPU.** A single MI325X loaded Llama 4 Scout (FP8) and accepted a 900,000-token prompt, with 4.96 million tokens of KV cache available on that one GPU. The first token took about nine minutes, so this demonstrates capacity, not production latency.[B4, B5]

This matters for coding agents working across repositories, research and RAG over large document sets, financial and legal analysis, and persistent customer-service agents: any workload where context accumulates.

06 · LARGE MODELS

405B, laid out four ways

Llama 3.1 405B in FP8 needs about 410 GB for weights alone. The H100 node can hold it only as one copy across all eight GPUs. The MI325X ran it as four 2-GPU copies, two 4-GPU copies, or one 8-GPU copy.[B2, B4]

Independent study by Bentaus. All data points measured solely by Bentaus.

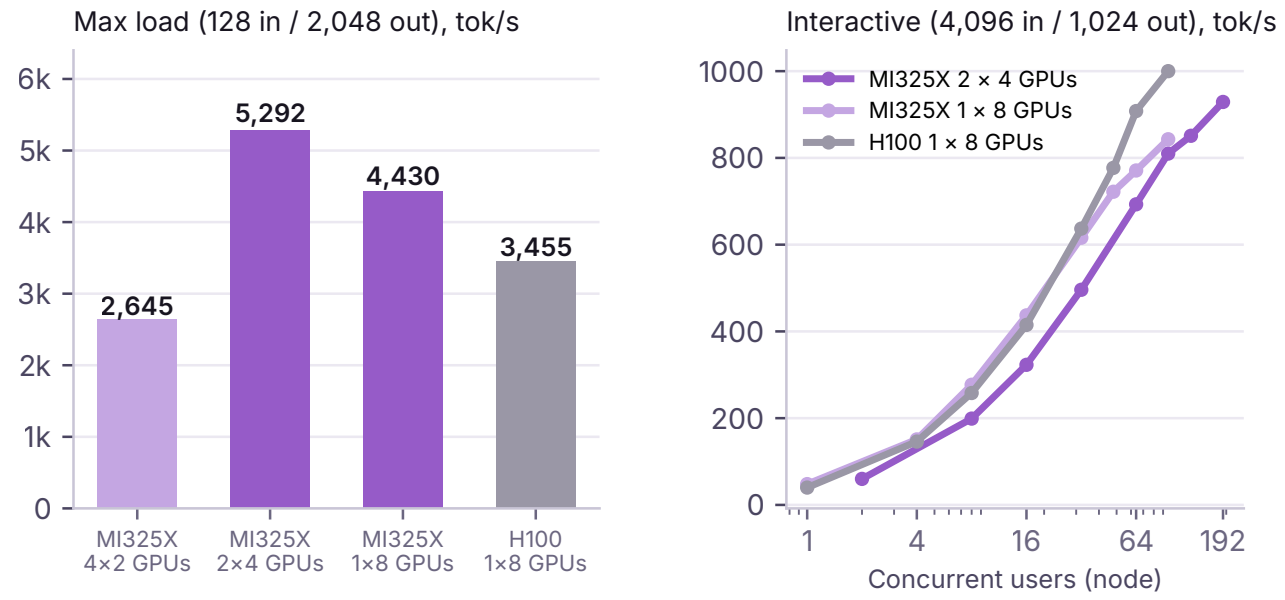

Figure 8. Llama 3.1 405B FP8. Left: maximum-load throughput by layout. Right: interactive throughput as users grow (4,096 in / 1,024 out). Source: B2, B4.

At maximum load, two 4-GPU copies produced **5,292 tok/s against 3,455: 1.53×, and 2.06× per dollar.** At interactive load the H100 is faster up to the 96 users it was tested at (1,000 vs 843 tok/s). The MI325X keeps adding users beyond that, to 929 tok/s at 192.

07 · WORKLOAD TRADEOFFS

Where the H100 performs better

On these workloads the H100 delivered higher performance in our tests:

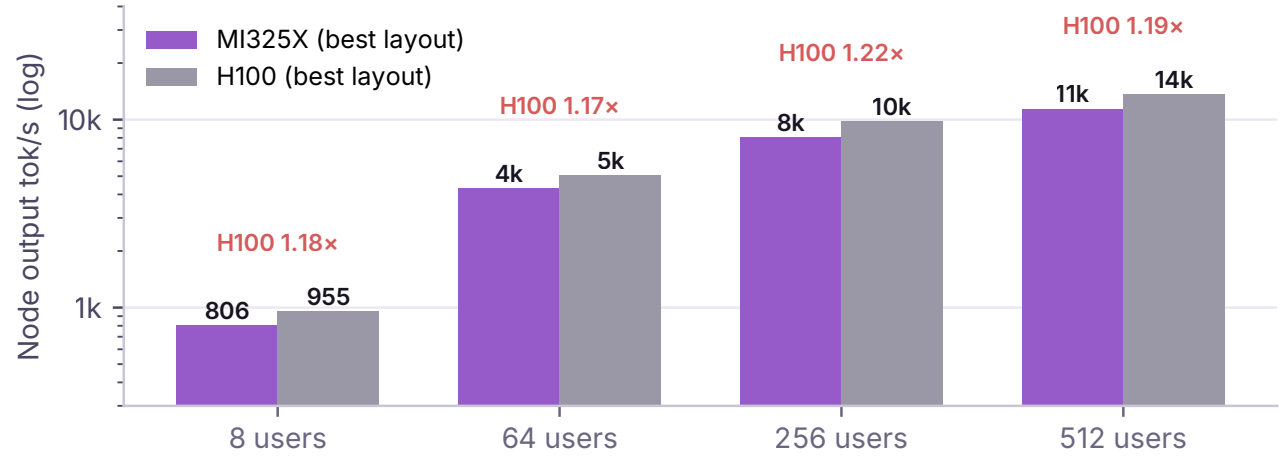

Figure 9. Llama 3.1 70B FP8 interactive serving (1,024 in / 1,024 out), best layout on each GPU at each user count. Source: B2, B4.

Independent study by Bentaus. All data points measured solely by Bentaus.

Llama 70B interactive serving

15–22% faster at 8 to 512 users. The MI325X is still ~13% ahead per dollar.

405B at 64–96 users

~18% faster on its 8-GPU layout. The MI325X is ~13% ahead per dollar.

Qwen3-30B FP8 user ramp

26% faster within the performance target, while AMD's MoE kernels are off. The MI325X is 7–10% ahead per dollar.

Energy per token

Lower in most uncapped tests: 0.34 vs 0.39 J on AMD's benchmark, about 2× at light load. The exception is max load with one copy per GPU (0.25 vs 0.27 J). Section 10 shows how Ziani narrows the gap.

Software configuration materially changes MI325X performance

With vLLM's defaults, 100K-token throughput was 59 tok/s; with AMD's AITER attention it was 150. Results depend on a validated ROCm, vLLM and AITER recipe, which Bentaus provides as part of the service.

08 · PRECISION

The MI325X is an FP8 machine

We compared Llama 70B at 16-bit (BF16), 8-bit (FP8) and two 4-bit formats (AWQ INT4 and MXFP4) on the same software, and scored quality on each with lm-evaluation-harness.[B4, O3]

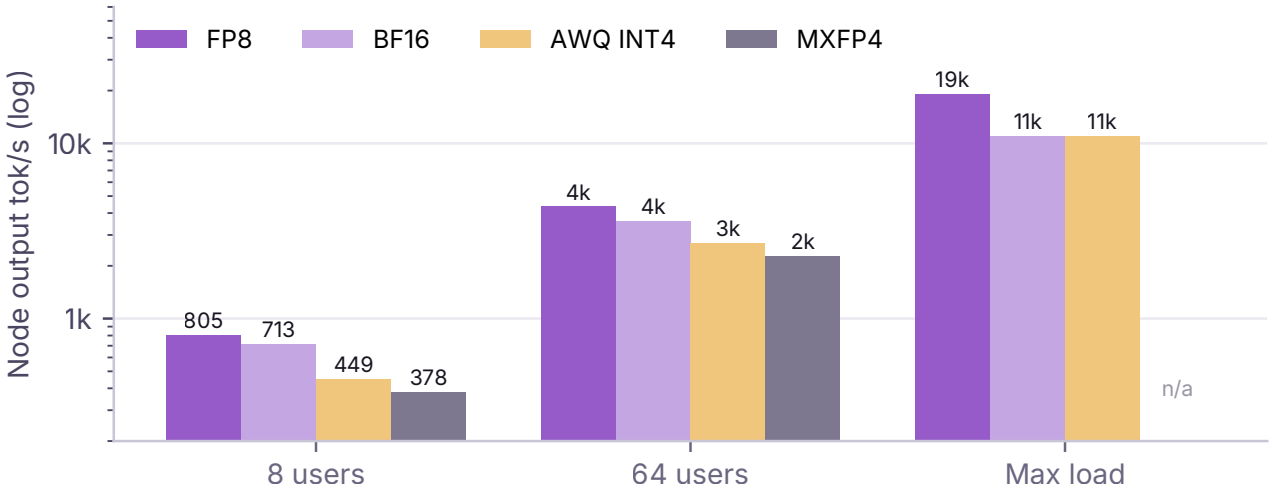

Figure 10. Node output throughput by precision and load, one copy on 8 MI325X GPUs. MXFP4 has no maximum-load result. Source: B1, B4.

Independent study by Bentaus. All data points measured solely by Bentaus.

Quality (higher is better)	FP8	BF16	AWQ INT4	MXFP4*
GSM8K (math)	92.6	93.0	93.1	92.0
HumanEval (coding)	81.7	79.9	78.7	65.2
MMLU-Pro (knowledge)	64.9	65.0	65.9	63.9

*AMD's MXFP4 checkpoint is built on Llama 3.3 70B; the others on Llama 3.1 70B.

FP8 was the fastest option at every tested load (1.73× BF16 at maximum load, 1.21× at 64 users), while maintaining comparable quality on the GSM8K, HumanEval and MMLU-Pro evaluations used in this study. Both 4-bit formats are slower than BF16, because the MI325X has no 4-bit math units and expands the weights on the fly. MXFP4 also lost about 15 points on coding. Native 4-bit belongs to the MI350 series, which Bentaus is testing next.

09 · SOFTWARE

Hardware matters. Software unlocks it.

The same MI325X hardware moved a lot with software. Our production stack added 14% over AMD's December 2025 image. AMD's AITER attention backend[A4] took 100K-token throughput from 59 to 150 tok/s, and more than doubled Qwen agent tasks within the performance target.[B1, B4, B6]

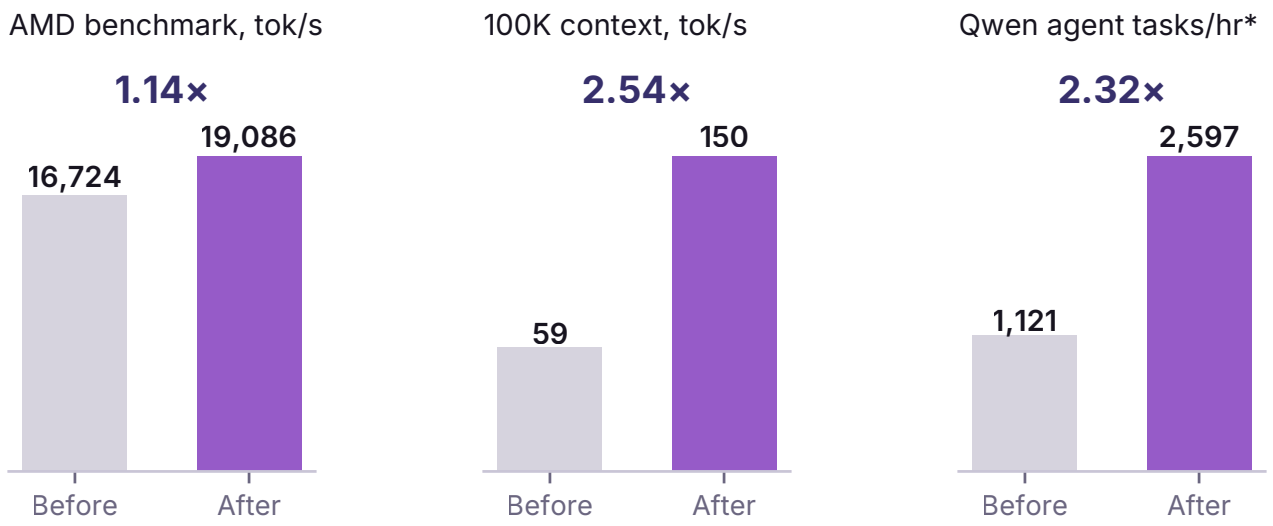

Figure 11. Same MI325X node before and after tuning. *Qwen3-30B BF16 agent tasks per hour at the highest load meeting the performance target, AITER attention off vs on. Source: B1, B4, B6.

The ROCm ecosystem is part of the product. ROCm, vLLM and AITER are where much of the MI325X's performance lives. Bentaus tunes the full stack, not just the server, and passes that through at the same $2.00 per GPU-hour.

Independent study by Bentaus. All data points measured solely by Bentaus.

10 · ZIANI

Maximum watts is not maximum useful work

Bentaus is an energy-first NeoCloud. Once Bentaus finds the right model, precision, layout and concurrency, **Ziani**, Bentaus' power-asset orchestration platform, sets and verifies the power each GPU draws while the workload runs. We stepped the MI325X power cap from 1,000 W down to 400 W, three runs per step.[B4]

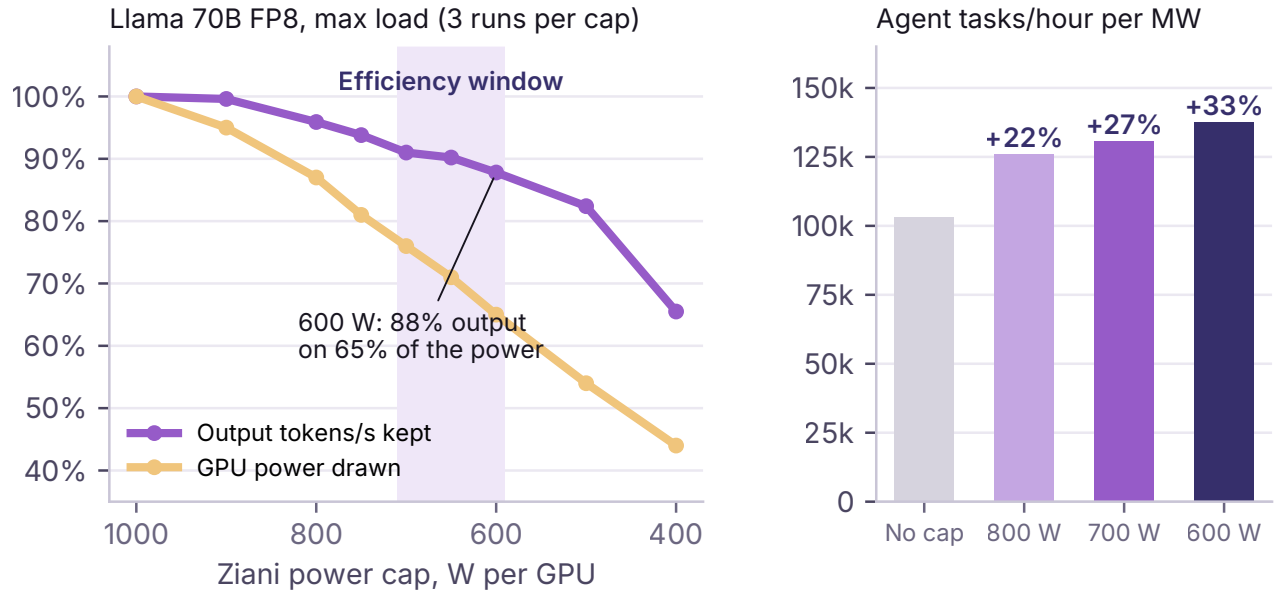

Figure 12. Left: Llama 70B FP8 at maximum load (served, one 8-GPU copy) under Ziani caps; spread across runs ≤1.2%. Right: Llama 70B agent tasks per hour per MW of GPU power, 8 agents per GPU, counting all completed tasks. Source: B4.

In our first power-cap run, Llama 70B kept **97% of its throughput on 23% less GPU power**, and energy per token fell 20% (1,503 → 1,452 tok/s; 7.27 → 5.58 kW per node at a 700 W cap).[B7] At maximum load, a 600 W cap kept **88% of output while drawing 65% of the power**, and energy per token fell 25% (0.51 → 0.38 J). On the agent workload the same cap kept 83% of tasks per hour on 62% of the power: **33% more completed tasks per MW.** The pattern held on 405B (96% of output at 800 W on 84% of the power) and at 100K context. The extra watts at the top of the MI325X's range buy little output.

Ziani changes power fast and predictably. Across 26 downward cap steps, the command returned in a median 7 ms, the new cap was confirmed in 9 ms, and power reached 90% of the change in a median 244 ms.[B4] At fleet scale, that creates the potential for flexible megawatts without stopping the workload.

> **The Bentaus metric stack.** Tokens per second · tokens per dollar · concurrent sessions per dollar · completed agent tasks per dollar · completed agent tasks per MW · flexible MW. The goal is not the most watts per GPU. It's the most useful work per megawatt.

Independent study by Bentaus. All data points measured solely by Bentaus.

11 · ECONOMICS

Tokens per dollar, test by test

At $2.00 per GPU-hour against $2.69 for the H100 (median of about 50 cloud listings), the MI325X delivered more output per dollar in every matched workload comparison reported in this study. The margin is narrowest where the H100 is faster per user, and widest where memory decides.

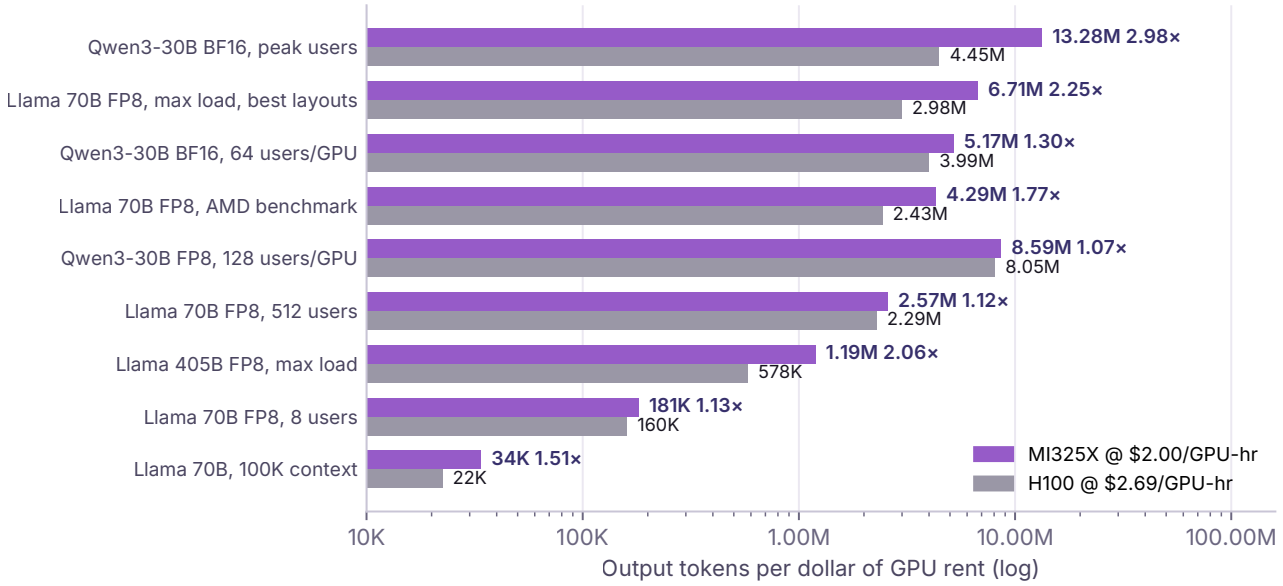

Figure 13. Output tokens per dollar of GPU rent: node tok/s × 3,600 ÷ node $/hr (MI325X $16.00, H100 $21.52). Ratios are MI325X ÷ H100. Agent tasks per dollar: Qwen BF16 2.18×, Qwen FP8 1.44×, Llama 70B 1.11×. Source: B2, B4.

The break-even test for newer GPUs. At $2.00 per GPU-hour, a GPU rented at $6.00 must deliver 3× the MI325X's useful output on the same workload just to tie on output per dollar. That's arithmetic on price, not a measurement of any other GPU.

Independent study by Bentaus. All data points measured solely by Bentaus.

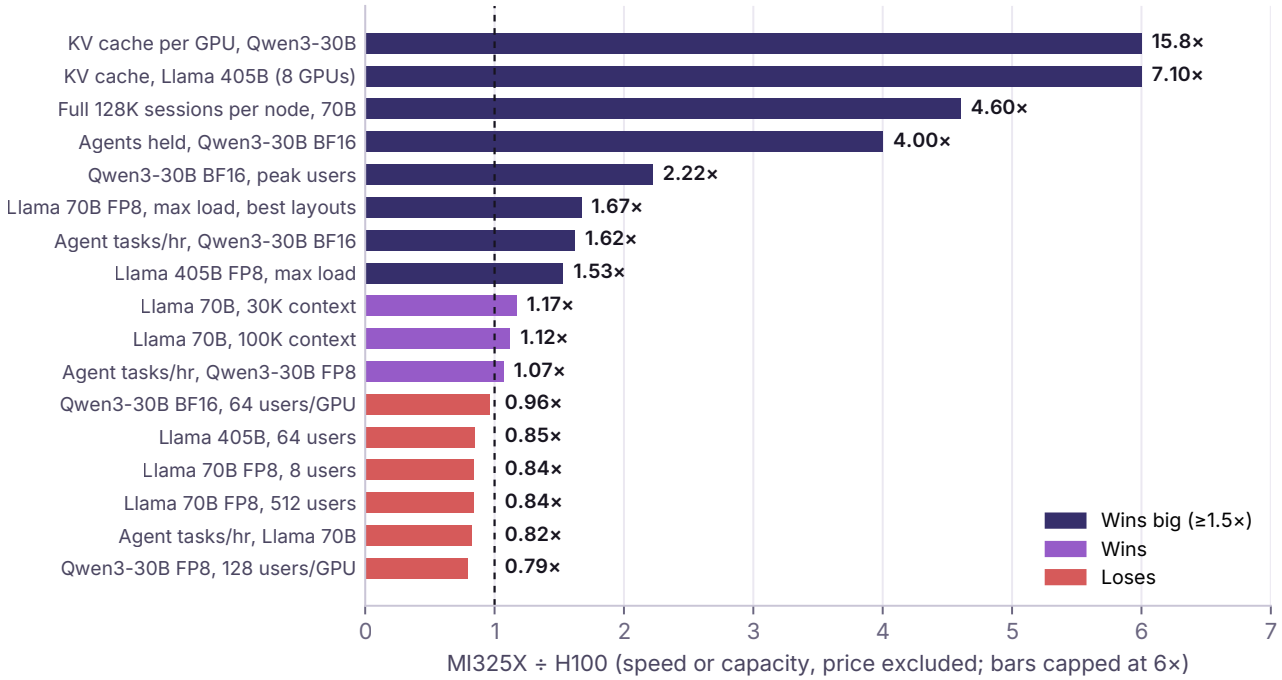

Figure 14. Speed and capacity scorecard, price excluded: MI325X ÷ H100 on each test. Capacity rows count what a node holds; the rest compare throughput. Source: B2, B4, B5.

Independent study by Bentaus. All data points measured solely by Bentaus.

12 · FIT

Where the MI325X Fits Best

Customer / use case	Why the MI325X fits	Measured proof	Metric to buy on
Best fit			
Agent platforms	Many persistent agents with growing context stay resident	4× agents within target (Qwen3-30B)	Agents/$, tasks/$, tasks/MW
Long-context and RAG	Large KV pools hold whole repositories and document sets	65 vs 14 full 128K sessions per node	Sessions per node
Large-model hosting	405B on 2 or 4 GPUs per copy; 70B on one GPU	5,292 vs 3,455 tok/s on 405B	Copies per node, tokens/$
High-concurrency serving	Keeps scaling after 80 GB parts fill their KV cache	59,026 vs 26,590 tok/s at peak	Aggregate tok/s, tokens/$
Strong economic fit			
Model-as-a-service	More copies and session state per node	8 vs 4 Llama 70B copies per node	Copies per GPU-hour
Batch and synthetic generation	Maximum throughput where first-token time doesn't matter	29,807 tok/s per node	Tokens/$
Unique Bentaus fit			
Power-constrained AI factories	Ziani trades a few percent of output for a large cut in power	+33% agent tasks per MW at 600 W	Tokens/MW, tasks/MW
Not the first choice			
Lightly loaded, latency-critical chat	The H100 delivered stronger per-user performance at lighter interactive loads.		
Native 4-bit (FP4) inference	The MI325X has no 4-bit math units; this belongs to the MI350 series.		

For lightly loaded, latency-sensitive inference, the H100 delivered stronger per-user performance in our tests. As concurrency, context, agents and model size grow, the MI325X becomes increasingly compelling.

Independent study by Bentaus. All data points measured solely by Bentaus.

13 · METHOD

Test setup

	AMD Instinct MI325X	NVIDIA H100
GPUs	8 × MI325X OAM, 256 GB HBM3E, 1,000 W[A1]	8 × H100 SXM, 80 GB HBM3, 700 W[N1]
Location	Bentaus us-lab, San Jose	Bentaus bv-lab, Netherlands
Software	vLLM 0.23.0, ROCm 7.14.1; AITER on for Llama, AITER attention on for Qwen agents	vLLM 0.23.0, CUDA (pinned vllm-openai image)
Models	amd/Llama-3.1-70B- and -405B-Instruct-FP8-KV[A5]; Llama 3.1 70B BF16, AWQ, MXFP4; Qwen3-30B-A3B-Instruct-2507 BF16 and FP8[O2]; Llama 4 Scout FP8. Same checkpoints on both GPUs.	
Price used	$2.00 / GPU-hr (Bentaus)	$2.69 / GPU-hr (median of ~50 cloud listings)
Performance target	Users: time to first token ≤2 s and inter-token latency ≤200 ms (P99). Agents: every call ≤3 s to first token and ≤200 ms per token, for ≥90% of tasks.	
Power	GPU board power, read every second. Default caps (1,000 W / 700 W) except in the Ziani sweeps.	

Bentaus production platform. Bentaus is deploying AMD Instinct MI325X at scale: 320 Supermicro-built 8-GPU nodes, 2,560 MI325X GPUs, on a DriveNets AI networking fabric. Supermicro's 8U MI325X platform[S1] pairs eight MI325X OAM accelerators with dual AMD EPYC™ processors over AMD Infinity Fabric™ links.

Limitations. One node per GPU type, at two sites. GPU board power, not wall power. Repeats were run on the AMD reproduction and all power-cap steps; most other points are single runs. Two MI325X GPUs ran at 98–100 °C and clocked ~5% lower, which likely understates the 8-GPU MI325X results by a few percent. AMD's MoE kernels were off for Qwen3-30B. On MI325X, four independent 2-GPU servers could not start together on this software stack. The Ziani maximum-load sweep used a served benchmark (1,200 requests), so its uncapped figure (14,313 tok/s) differs from the offline benchmark. Prices are inputs and change per-dollar results proportionally.

Independent study by Bentaus. All data points measured solely by Bentaus.

REFERENCES

Sources

BENTAUS PRIMARY MEASUREMENTS

[B1] Bentaus, "MI325X: reproducing AMD's vLLM benchmark," Bentaus MI325X validation test report, 2 Oct 2026.

[B2] Bentaus, "H100: test results," Bentaus MI325X validation test report, 2 Oct 2026.

[B3] Bentaus, "MI325X vs H100," Bentaus MI325X validation comparison report, 2 Oct 2026.

[B4] Bentaus, "MI325X: test results," Bentaus MI325X validation test report, 3 Oct 2026.

[B5] Bentaus, MI325X Memory Showcase results (KV capacity by layout), Oct 2026.

[B7] Bentaus, "First results: MI325 and MI300X," MI325X power-cap validation report, 1 Oct 2026.

[B6] Bentaus, MI325X validation raw measurements, MI325X and H100 (Raw_Data_MI325X_H100, issued 4 Oct 2026).

AMD

[A1] AMD, AMD Instinct™ MI325X Accelerator data sheet. amd.com/content/dam/amd/en/documents/instinct-tech-docs/product-briefs/instinct-mi325x-datasheet.pdf

[A2] AMD, Performance results with AMD ROCm™ software — vLLM, MI300X platform. amd.com/en/developer/resources/rocm-hub/dev-ai/performance-results.html

[A3] AMD ROCm, MAD (Model Automation and Dashboarding) benchmark scripts, commit 8294c5cac7. github.com/ROCm/MAD

[A4] AMD ROCm, AITER — AI Tensor Engine for ROCm. github.com/ROCm/aiter

[A5] AMD, amd/Llama-3.1-70B-Instruct-FP8-KV and amd/Llama-3.1-405B-Instruct-FP8-KV. huggingface.co/amd

[A6] AMD ROCm, vLLM container image rocm/vllm:rocm7.0.0_vllm_0.11.2_20251210. hub.docker.com/r/rocm/vllm

PLATFORM AND COMPARISON HARDWARE

[S1] Supermicro, GPU A+ Server AS-8126GS-TNMR — 8U system with AMD Instinct™ MI325X 8-GPU. supermicro.com/en/products/system/gpu/8u/as-8126gs-tnmr

[N1] NVIDIA, H100 Tensor Core GPU specifications. nvidia.com/en-us/data-center/h100/

SOFTWARE AND MODELS

[O1] vLLM project. github.com/vllm-project/vllm

[O2] Qwen, Qwen3-30B-A3B-Instruct-2507. huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507

[O3] EleutherAI, lm-evaluation-harness. github.com/EleutherAI/lm-evaluation-harness

AMD, AMD Instinct, ROCm, EPYC, Infinity Fabric and combinations thereof are trademarks of Advanced Micro Devices, Inc. NVIDIA and H100 are trademarks of NVIDIA Corporation. Supermicro is a trademark of Super Micro Computer, Inc. Results are Bentaus measurements on the configurations described; performance varies with model, software, settings and workload.

Disclaimer: Independent study conducted by Bentaus. All data points were measured solely by Bentaus. AMD did not sponsor, commission or fund this study.

BENTAUS NEOCLOUD · ENERGY-FIRST

Built for workloads that get bigger over time: more users, more context, more agents and larger models.

The fastest GPU isn't always the most productive infrastructure.

Bring us your model and your workload. Bentaus returns its profile on AMD Instinct: the right precision, layout and concurrency, the tokens and tasks per dollar, and, with Ziani, the power it really needs. We tune the software around the accelerator, and the energy around the workload.

Built for throughput
29,807 tok/s per 8-GPU node

Built for memory
256 GB HBM3E per GPU

Built for concurrency
59,026 tok/s at peak Qwen load

Built for agents
4× the agents within target

Built for large models
5,292 tok/s on Llama 405B, 1.53× the H100

Built for energy efficiency
+33% agent tasks per MW with Ziani

Bentaus