Bentaus

BENTAUS NEOCLOUD · EXECUTIVE BRIEF · OCTOBER 2026

Where the AMD Instinct™ MI325X Wins

We ran the MI325X and the NVIDIA H100 through the same production workloads on the Bentaus NeoCloud. On common benchmarks both performed well. On the workloads where AI demand is heading (many users, multi-step agents, long context and large models) the MI325X pulled ahead, and it's the GPU a standard benchmark most underestimates.

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

bentaus.com

Running the Ziani Power Asset Orchestrator

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

01 · THE THESIS

The workload decides the winner

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.

At lower concurrency, the H100 often delivered higher per-user performance. As users, agents and context increased, the MI325X's larger HBM capacity allowed the workload to continue scaling. Our Qwen3-30B user ramp shows the crossover:

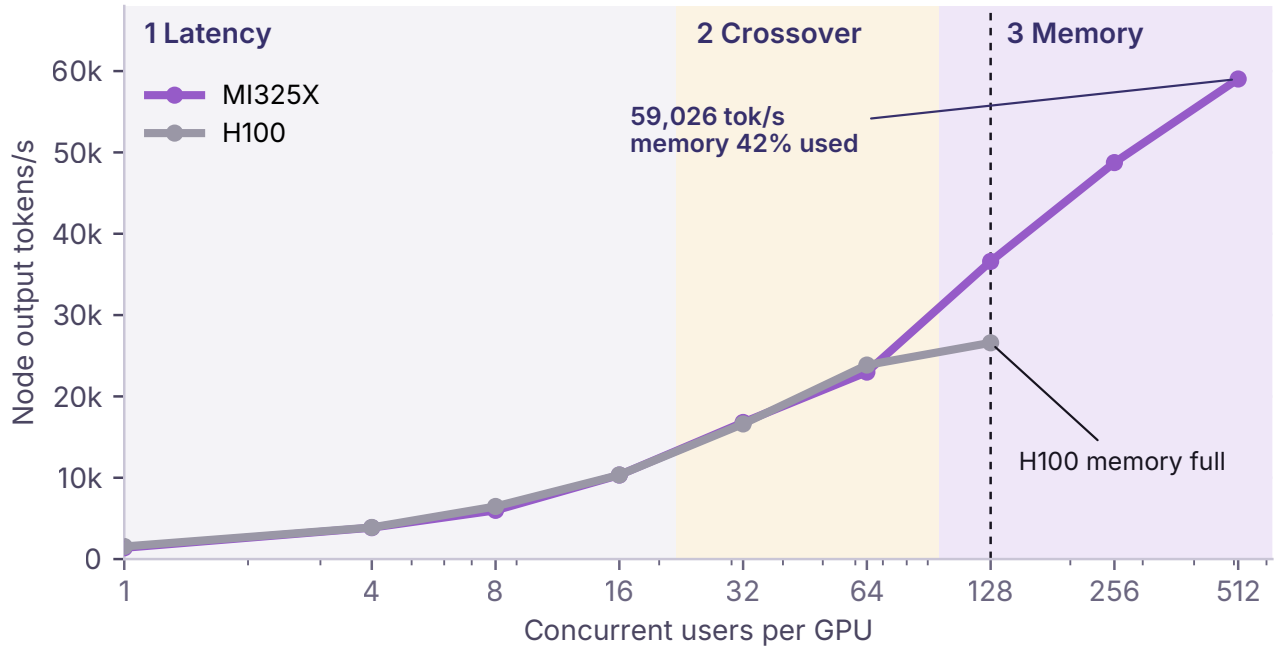

Figure 1. Node output throughput as concurrent users grow, Qwen3-30B-A3B BF16, one copy per GPU, same software on both. Both met the performance target (first token ≤2 s, inter-token ≤200 ms, P99) up to 64 users per GPU.

1 · Latency zone

Few users. The H100 matches or slightly leads.

2 · Crossover zone

Speed is tied; MI325X pricing gives 1.3× the tokens per dollar.

3 · Memory zone

The H100's memory is full at 128 users per GPU (26,590 tok/s). The MI325X climbs to **59,026 tok/s, 2.22× the H100**, with 42% of its KV cache used.

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

02 · MEMORY

Memory changes what you can run

Configuration	MI325X (256 GB)	H100 (80 GB)
Llama 3.1 70B FP8, one copy per GPU	29,807 tok/s per node	Does not fit with room for users
Llama 3.1 405B FP8, two 4-GPU copies	5,292 tok/s per node	Does not fit · 320 GB < 410 GB
Full 128K-token sessions per node	65	14
Llama 4 Scout, 900K-token prompt, one GPU	Runs · 4.96M KV tokens	Weights alone exceed 80 GB

Agents: completed work, not just tokens

Each agent task is 15 model calls with context growing to ~60,000 tokens. A load level counts only if at least 90% of tasks meet the performance target. On Qwen3-30B, the MI325X held **64 agents per node against 16** and completed **2,597 tasks per hour against 1,600: 1.62× the work and 2.18× the tasks per dollar.** At FP8, 2,594 vs 2,427 tasks per hour. On Llama 70B agents the H100 was slightly faster.

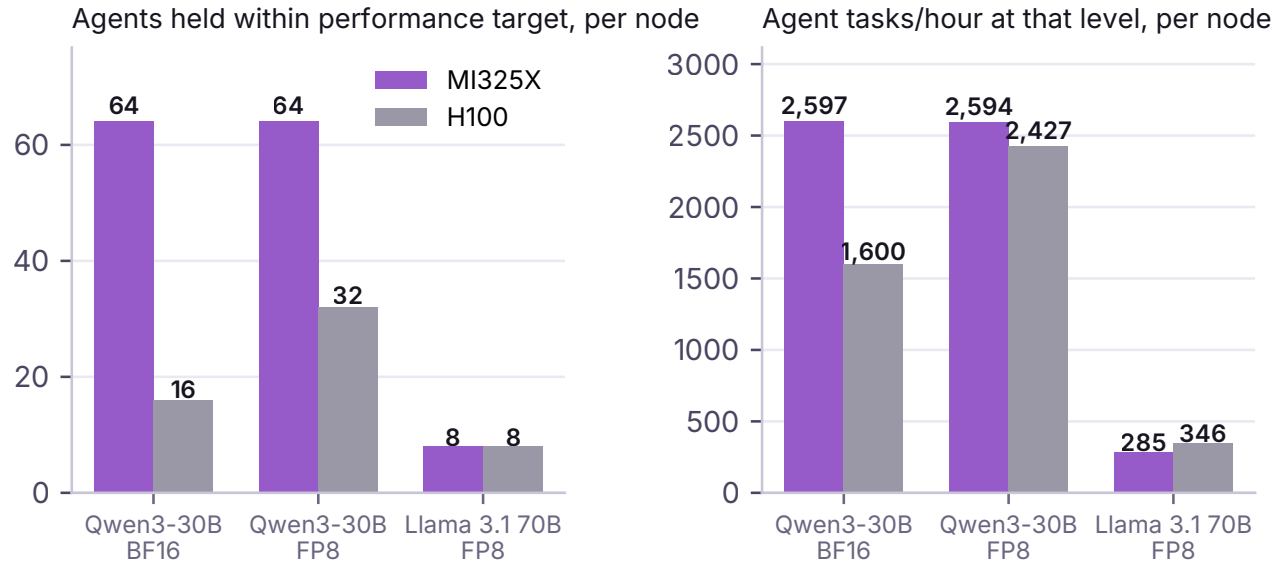

Figure 2. Highest agent load per node meeting the performance target, and tasks per hour at that load. MI325X used AMD's AITER attention backend.

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

03 · ECONOMICS

What a dollar buys

At $2.00 per GPU-hour on the Bentaus NeoCloud against $2.69 for the H100 (median of ~50 cloud listings), the MI325X delivered more output per dollar in every matched workload comparison reported in this study, from 1.07× to 3×. It costs less per hour and holds more per node.

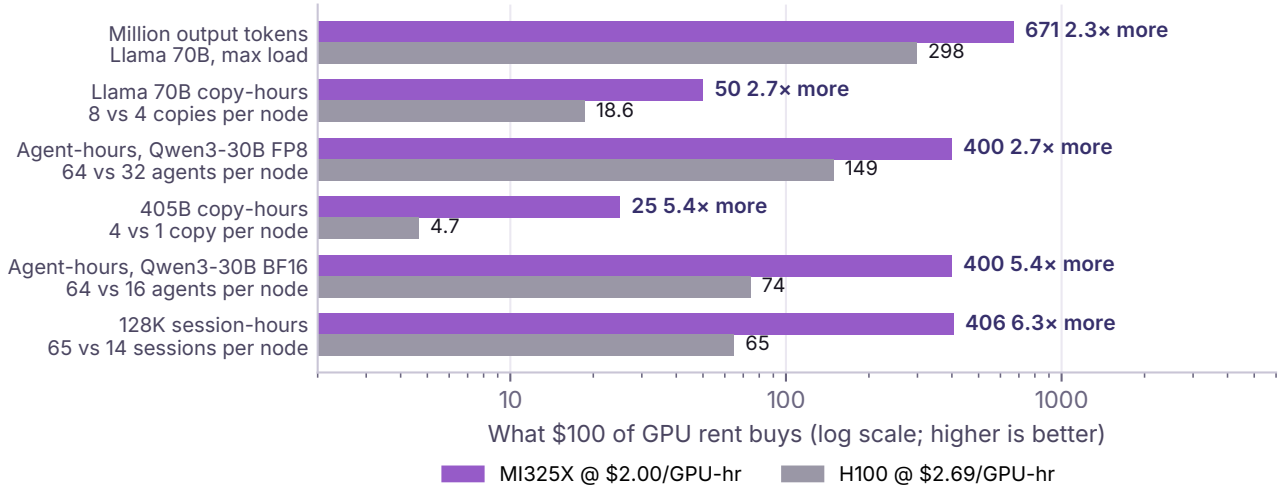

Figure 3. What $100 of GPU rent buys, from measured capacity per node (MI325X node $16.00/hr, H100 node $21.52/hr).

Definitions. An agent-hour is one agent actively running multi-step tasks for an hour at the highest load meeting the performance target. A session-hour is one full 128K-token context resident in GPU memory for an hour. A copy-hour is one complete model replica serving for an hour.

Where the H100 performs better

Interactive Llama 70B serving

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

Energy per token, uncapped

Lower in most tests, which is where Ziani comes in.

405B at 64–96 users

~18% faster on its only layout.

Qwen3-30B FP8 user ramp

26% faster within target; MI325X 7–10% ahead per dollar.

Test basis. One 8-GPU node of each type: 8 × MI325X (256 GB HBM3E, 1,000 W) and 8 × H100 SXM (80 GB, 700 W), vLLM 0.23.0 on ROCm 7.14.1 and CUDA, same model checkpoints, prompts and ramps. GPU board power, not wall power. Prices: MI325X $2.00, H100 $2.69 per GPU-hour. Full method, limitations and sources: *Where the AMD Instinct™ MI325X Wins*, Bentaus white paper, October 2026. AMD, AMD Instinct and ROCm are trademarks of Advanced Micro Devices, Inc. NVIDIA and H100 are trademarks of NVIDIA Corporation.

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

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

04 · ZIANI

Maximum watts is not maximum useful work

Once Bentaus finds the right model, precision, layout and concurrency for a workload, **Ziani**, Bentaus' power-asset orchestration platform, sets and verifies the power each GPU draws while it runs. The extra watts at the top of the MI325X's range buy little output.

97%
of Llama 70B output on 23% less GPU power, energy per token −20%

+33%
agent tasks per MW at a 600 W cap

244 ms
median for power to reach 90% of a cap change; command returns in 7 ms

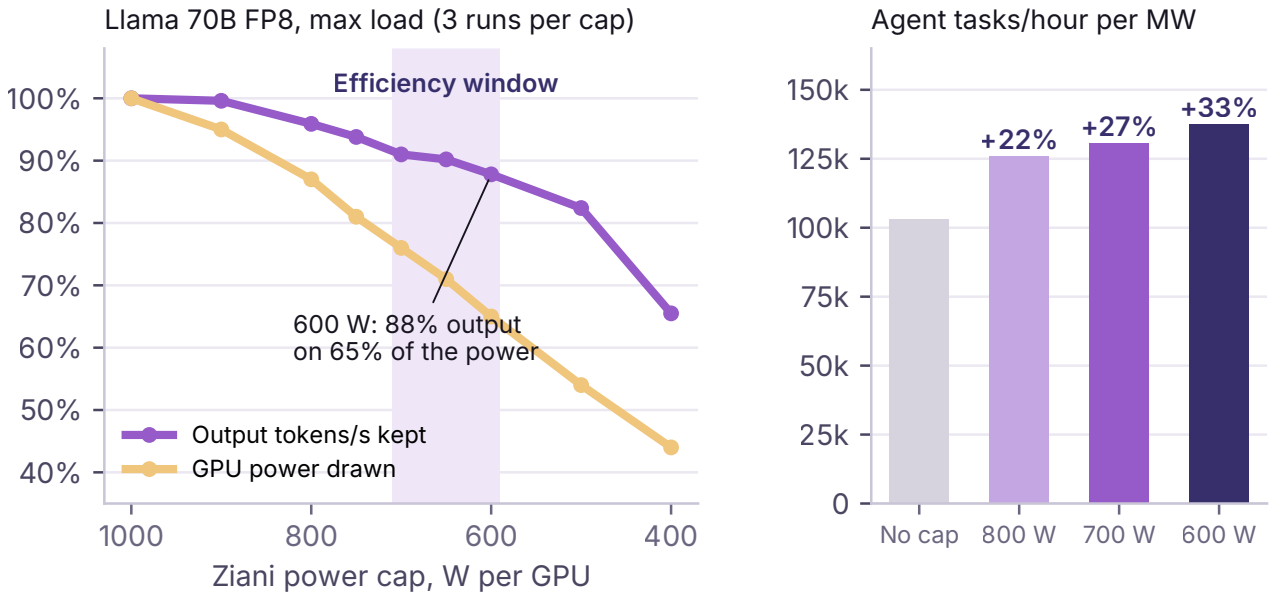

Figure 4. Left: Llama 70B FP8 at maximum load under Ziani caps, 3 runs per cap; 600 W kept 88% of output on 65% of the power. Right: agent tasks per hour per MW of GPU power.

> **The goal is not the most watts per GPU. It's the most useful work per megawatt.** Tokens per second · tokens per dollar · agents per dollar · completed tasks per MW · and, at fleet scale, the potential for flexible MW without stopping the workload.

Software is part of the product

On the same hardware, Bentaus' production stack ran AMD's benchmark 14% faster than AMD's December 2025 image (19,086 vs 16,724 tok/s). AMD's AITER attention backend took 100K-token throughput from 59 to 150 tok/s and more than doubled Qwen agent tasks within target. Bentaus tunes ROCm, vLLM and AITER as part of the service.

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

05 · FIT

Where the MI325X Fits Best

Customer / use case	Measured proof	Metric to buy on
Best fit		
Agent platforms	4× agents within target; 1.62× tasks per hour	Agents/$, tasks/$, tasks/MW
Long-context and RAG	65 vs 14 full 128K sessions per node	Sessions per node
Large-model hosting	5,292 vs 3,455 tok/s on 405B; 70B on one GPU	Copies per node, tokens/$
High-concurrency serving	59,026 vs 26,590 tok/s at peak	Aggregate tok/s, tokens/$
Strong economic fit		
Model-as-a-service	8 vs 4 Llama 70B copies per node	Copies per GPU-hour
Batch and synthetic generation	29,807 tok/s per node	Tokens/$
Unique Bentaus fit		
Power-constrained AI factories	+33% agent tasks per MW with Ziani	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	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.

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