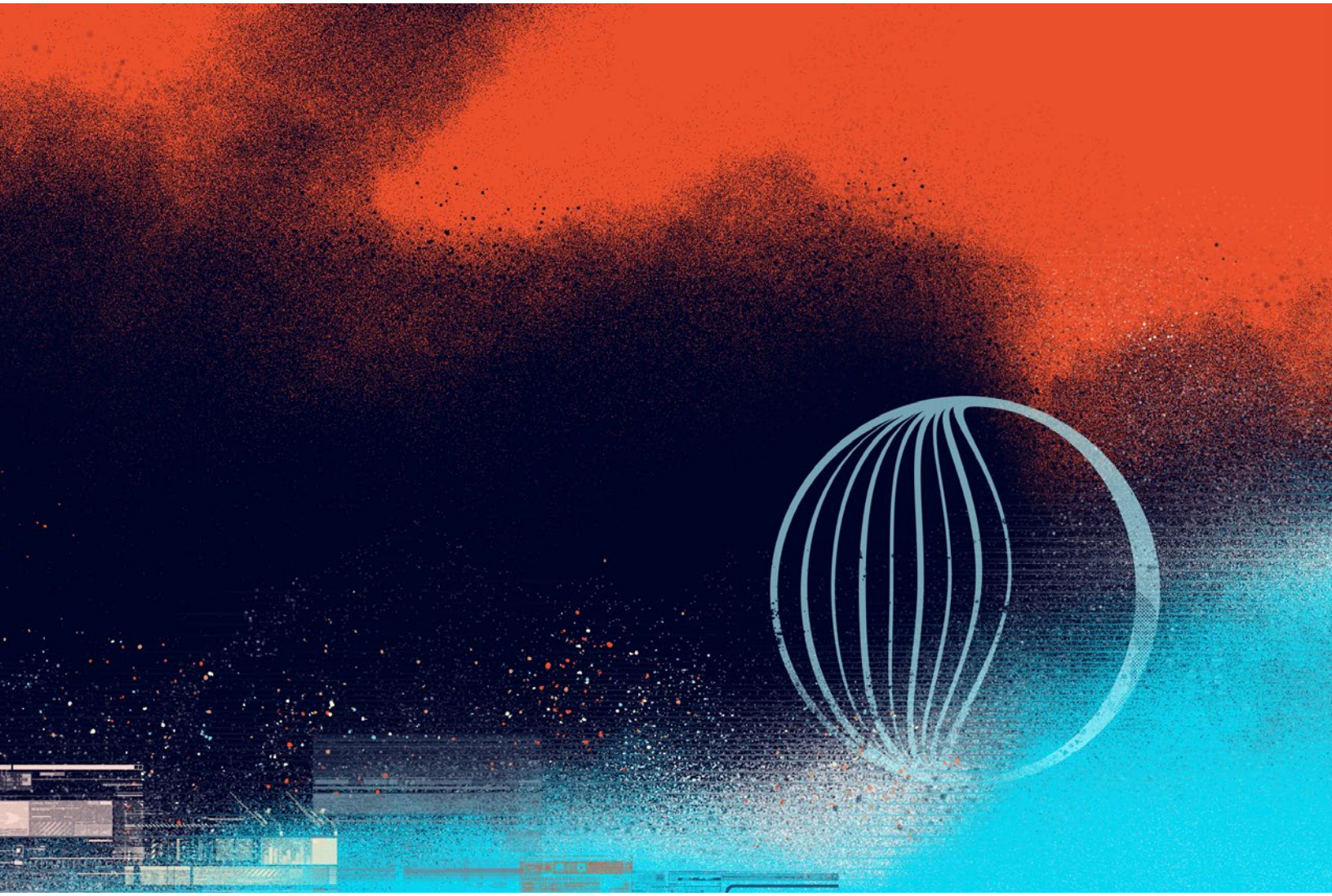
signal65



EXECUTIVE SUMMARY

AMD Instinct MI355X GPU platform vs NVIDIA HGX B200 GPU platform: A Three-Level Inference Infrastructure Evaluation

AUTHORS

Mitch Lewis
Performance Analyst | Signal65

Russ Fellows
VP, Labs | Signal65

Cameron Moccari
Operations Director | Signal65

IN PARTNERSHIP WITH

AMD
together we advance_

JULY 2026

Overview

Traditional approaches to evaluating AI infrastructure often rely solely on raw performance metrics. While performance establishes a strong baseline, enterprises must also weigh a broader range of economic and business metrics to make fully informed infrastructure decisions.

In this study, Signal65 evaluated AMD Instinct MI355X against NVIDIA HGX B200 across three lenses — raw throughput, cost per token, and translated business outcomes — using three production models (GPT-OSS-120B, Qwen3-Next-80B, Kimi-K2.6) on cloud-hosted 8-GPU nodes. This study demonstrates that infrastructure economics, not raw silicon speed alone, determine real business value.

Notably, MI355X was found to achieve significant advantages across all three levels of evaluation – achieving higher performance for 2 of the 3 models tested, lower cost per token, and advantages across three distinct business workloads.

1.3x
More financial docs/hour

53%
Lower cost per document

2.15x
More tokens per dollar — best case across 3 models tested

39.6%
Lower Cost

42%
Lower p99 latency

Same Interactive Users Served

Raw Performance

When evaluating raw performance, MI355X led on 2 of 3 models tested. For the one exception, Kimi-K2.6, NVIDIA B200 outperformed MI355X by a median of 17%. At high concurrency and long context, however, MI355X reverses this and leads by up to 1.96x higher throughput due to its larger HBM capacity.

Model	Median Throughput Ratio (MI355X/B200)	Range	Winner
GPT-OSS-120B (ATOM vs. TRT-LLM)	**1.30x**	1.10–2.25x	**AMD MI355X**
Qwen3-Next-80B (vLLM)	**1.27x**	1.06–2.54x	**AMD MI355X**
Kimi-K2.6 (vLLM)	**0.86x** (B200 +17%)	0.76–1.96x	**NVIDIA B200**

Cost & Price Performance

The cost of each platform was evaluated using blended on-demand pricing across cloud providers. Average pricing was calculated at $4.71/GPU-hr for MI355X and $7.80/GPU-hr for B200 — a 39.6% infrastructure cost advantage for MI355X on its own, before performance is even factored in.

Combined with performance, this cost advantage translates into significantly more tokens per dollar for MI355X across every model tested.

Model	Median Tokens/$ Advantage
GPT-OSS-120B	2.15× MI355X
Qwen3-Next-80B	2.11× MI355X
Kimi-K2.6	1.42× MI355X

When evaluating price-performance, MI355X achieves advantages across all three models tested. Most notably, MI355X regains a significant advantage on Kimi-K2.6, the one model in which NVIDIA B200 achieved higher raw performance, with 1.42x more tokens per dollar.

Business Workload Impact

Beyond raw performance and cost, both platforms were evaluated against three representative business workloads – Financial Analysis, Long Report Generation, and an Interactive RAG Chatbot. Each workload was profiled for realistic token, concurrency, and latency requirements and mapped to matching test configurations.

Batch — Financial Analysis

- MI355X achieves 1.3× more documents processed per hour, at 53% lower cost per document (holds across 64–256 concurrency).

Batch — Long Report Generation

- MI355X trades ~5% raw throughput for 37% lower cost per page at representative concurrency; at the highest concurrencies tested, MI355X leads on both throughput and cost (up to 69% lower cost/page).

Interactive — RAG Chat Application

- At equal concurrency, MI355X delivers the same served-user population at 42–44% lower p99 latency and 39.6% lower cost per hour.
- MI355X sustains roughly 2× the concurrent user load before breaching the latency SLA, giving enterprises headroom to support more users or absorb usage surges.

Conclusion

Making well informed decisions about AI infrastructure depends on more than raw performance alone. Enterprises must additionally consider how the performance and cost characteristics of a platform intersect to support their required AI workloads.

This study evaluated AMD Instinct MI355X and NVIDIA HGX B200 across raw performance, cost, and three representative business workloads. Across all three dimensions, MI355X delivers a consistent advantage — even in the one instance where NVIDIA is faster in raw throughput. For enterprises deploying AI applications, this points to infrastructure economics, not silicon speed alone, as the deciding factor.

Full methodology, detailed configuration tables, and per-run data are available in the complete Signal65 report.

Important Information About this Report

CONTRIBUTORS

Mitch Lewis
Performance Analyst | Signal65

Russ Fellows
VP, Labs | Signal65

Cameron Moccari
Operations Director | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

IN PARTNERSHIP WITH

AMD
together we advance_

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION

Signal65 I signal65.com