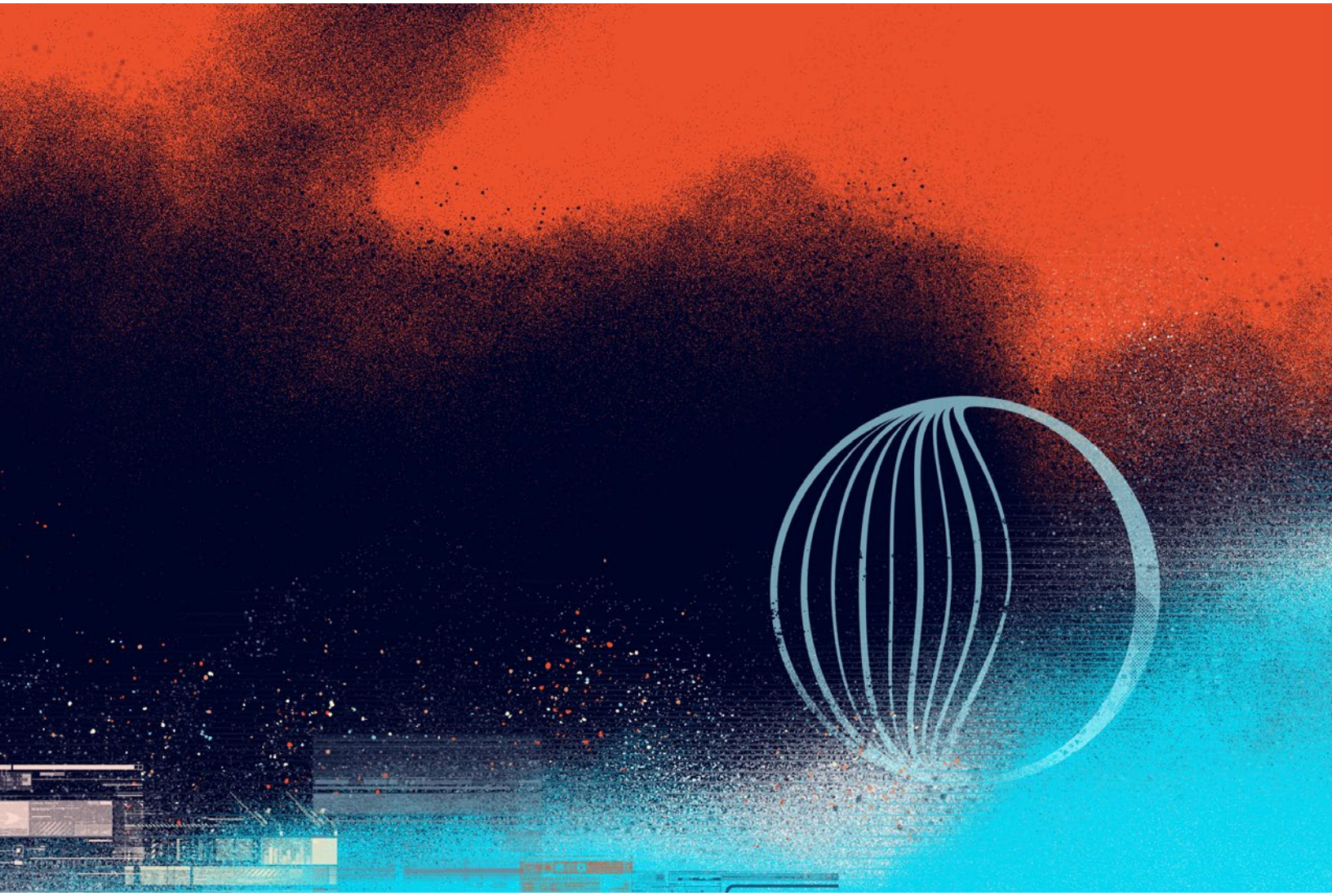
signal65



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

JULY 2026

IN PARTNERSHIP WITH

AMD
together we advance_

Executive Summary

When enterprises evaluate AI infrastructure, traditional approaches often focus primarily on raw performance metrics. While important, assessing hardware solely through maximum token throughput delivers an incomplete view of operational reality. Similarly, cost examined in isolation tells only part of the story. Neither alone provides the data organizations need to make informed infrastructure decisions, leaving a gap in understanding how infrastructure performs for real business objectives.

To bridge this gap, Signal65 conducted a three level evaluation of AMD Instinct MI355X GPU platform compared to NVIDIA HGX B200 GPU platform, evaluating three aspects:

- **Raw Hardware Performance:** Measured total token throughput (tokens per second) across each vendor's optimized software inference stack.
- **Operational Economics:** The cost-efficiency of that performance calculated using current market GPU-hour hosting rates.
- **Business Workload Impact:** The translation of hardware capacity and economics into tangible enterprise outcomes, evaluating both batch processes and latency sensitive interactive workloads.

Key Takeaways

- 1.3× more financial analysis documents per hour at **53% lower cost per document**
- **Up to 2.15× more tokens per dollar —** across all three models tested, including where B200 leads on raw throughput
- Same interactive user population served at **39.6% lower cost** with 42% lower p99 latency

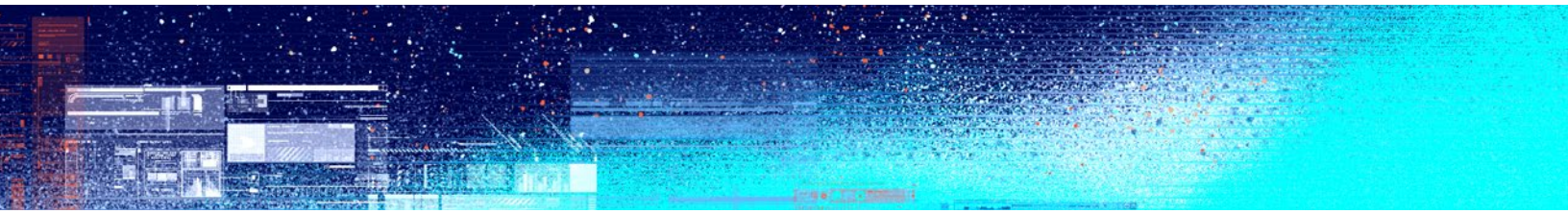
The testing encompassed three production-class large language models: GPT-OSS-120B, Qwen3-Next-80B, and Kimi-K2.6. These models were evaluated across various input/output context sizes and request concurrency levels on standard 8-GPU nodes:

Key Findings:

- AMD MI355X delivered superior business outcomes compared to NVIDIA B200 across three distinct workload configurations — batch document processing, long-form generation, and latency-sensitive interactive serving — each mapping to a real application scenario.
  - **Financial Analysis:** 1.3x more documents processed per hour, with 53% lower cost per document
  - **Long Report Generation:** Trades ~5% throughput for 37% lower cost per page vs NVIDIA B200
  - **Interactive RAG Chat Application:** serves the same number of users at 39.6% lower cost per hour and 42% lower p99 latency.

- AMD MI355X achieved higher performance than NVIDIA B200 for 2 out of the 3 models tested:
  - 1.3x greater throughput when running GPT-OSS-120B.
  - 1.27x greater throughput than B200 when running Qwen3-Next-80B.
  - NVIDIA B200 delivered 17% higher throughput than MI355X when running Kimi-K2.6, while MI355X leads on price-performance
- MI355X achieves more value per dollar:
  - GPT-OSS-120B – 2.15x more tokens per dollar
  - Kimi-K2.6 – 1.42x more tokens per dollar
  - Qwen3-Next-80B: 2.11x more tokens per dollar

Ultimately, MI355X delivers consistent economic and workload advantages across all three models tested, including instances where NVIDIA outperforms on raw throughput. This demonstrates that infrastructure cost and workload efficiency are just as consequential as raw performance in determining real-world business value.



Test Methodology and Infrastructure Setup

> To ensure a fair comparison, both hardware platforms were evaluated utilizing like-for-like software configurations on equivalent 8-GPU server nodes. AMD Instinct MI355X GPUs used in this evaluation were provided by TensorWave, an AMD-focused cloud provider cloud provider offering large-scale GPU infrastructure for AI training and inference.

Hardware & Hosting Environments

- **AMD Instinct MI355X GPU Platform:** 8-GPU node hosted via TensorWave
- **NVIDIA HGX B200 GPU Platform:** 8-GPU node hosted via RunPod

Software Inference Stacks

Models were evaluated using both vendor-native inference stacks — AMD ATOM and NVIDIA TensorRT-LLM — and the open-source vLLM framework. GPT-OSS-120B was evaluated using each vendor's native stack, while Qwen3-Next-80B and Kimi-K2.6 were evaluated using vLLM on both platforms, providing a consistent like-for-like software environment for those models.

Model	AMD MI355X Stack	NVIDIA B200 Stack	Selection Rationale
GPT-OSS-120B	AMD ATOM stack	NVIDIA TensorRT-LLM	TensorRT-LLM represents NVIDIA's fastest stack for this specific model.
Qwen3-Next-80B	AMD vLLM stack	NVIDIA vLLM stack	vLLM is the common inference stack for both platforms, giving a like-for-like software comparison
Kimi-K2.6	AMD vLLM stack	NVIDIA vLLM stack	vLLM represents the fastest common execution environment for both vendors.

Figure 1: *Software Stack Configurations*

Signal65 Comment: *In our testing TRT-LLM measured 1.1x to 2.0x lower throughput than vLLM for Qwen3-Next-80B on B200. AMD ATOM showed marginally higher throughput than vLLM on MI355X, but for Qwen3-Next-80B and Kimi-K2.6 both platforms were standardized on vLLM to ensure a consistent like-for-like comparison; GPT-OSS-120B used each vendor's native stack (ATOM vs. TensorRT-LLM).*

Key Performance Metrics

Tensor Parallelism Configurations

Tensor parallelism (TP) width determines how many GPUs are used per model instance. Configurations varied by model based on scale and testing scope:

- GPT-OSS-120B: Evaluated at TP=1, as the model's size and low active-parameter count make it well suited for single-GPU MoE serving
- Qwen3-Next-80B: Evaluated across TP=1, TP=2, and TP=4 to capture performance across a range of deployment configurations
- Kimi-K2.6: Evaluated at TP=4 due to its parameter scale requiring multi-GPU deployment

Throughput Metric Normalization: Reported throughput figures are calculated as per-GPU total (input + output) tokens per second. This value is derived by dividing the aggregate endpoint throughput by the active tensor-parallel width, enabling linear comparison across diverse topologies.

Pricing Methodology

Operational economics are driven by the current market pricing for on-demand cloud resource allocation. The prices include both cloud providers utilized during testing, as well as two additional providers that offer both MI355X and B200 instances[1]. Providers were selected to span both ends of the pricing spectrum, ensuring the blended average reflects realistic market pricing rather than either best-case or worst-case scenarios. It should be noted, however, that cloud pricing can vary due to several factors including overall demand and any negotiated discounts.

[1] All price data collected as of July 10th 2026.

Cloud Provider	AMD MI355X Price ($/GPU/Hr)	NVIDIA B200 Price ($/GPU/Hr)
Vultr	$2.59	$3.50
TensorWave	$2.95	-
RunPod	-	$5.89
Oracle	$8.60	$14.00
Average	**$4.71**	**$7.80**

Figure 2: *Cloud Pricing*

Signal65 Comment: *The MI355X is notably* ***39.6% lower per GPU-hour*** *on average compared to NVIDIA B200. This gap fundamentally shapes all cost-efficiency metrics presented in this report. In our view, this pricing differential is significant relative to the highly competitive performance observed between the two platforms, and enterprises should weigh it carefully when evaluating total infrastructure cost.*

Raw Hardware Performance Analysis

The following results reflect inference throughput measured on 8-GPU nodes, normalized per GPU to enable consistent comparison across configurations.

Total Per-GPU Token Throughput Summary

The table below displays the median ratios and the comprehensive ranges observed during testing across the complete sweep of request concurrency and prompt shapes. Ratios represent MI355X throughput divided by B200 throughput. A ratio above 1.0 indicates MI355X leads; below 1.0 indicates B200 leads.

Model Evaluation Configuration	Median Throughput Ratio (MI/B)	Observed Range	Performance Winner
GPT-OSS-120B (ATOM vs. TRT-LLM, TP=1)	**1.30x**	1.10–2.25x	**AMD MI355X**
Qwen3-Next-80B (vLLM vs. vLLM, TP=1 - 4)	**1.27x**	1.06–2.54x	**AMD MI355X**
Kimi-K2.6 (vLLM vs. vLLM, TP=4)	**0.86x** (B200 leads at 1.17x)	0.76–1.96x	**NVIDIA B200**

Figure 3: *Performance Summary*

Model Observations

GPT-OSS-120B

AMD MI355X leads across every tested prompt configuration and concurrency level. As the context length expands and request concurrency increases, AMD's architectural advantage grows significantly. Under standard token shapes, the advantage remains near the median 1.3x advantage, but it extends to 2.25x the throughput of the B200 during long-context operations. This scaling behavior reflects MI355X's larger HBM capacity, which provides additional headroom as memory pressure increases.

Qwen3-Next-80B

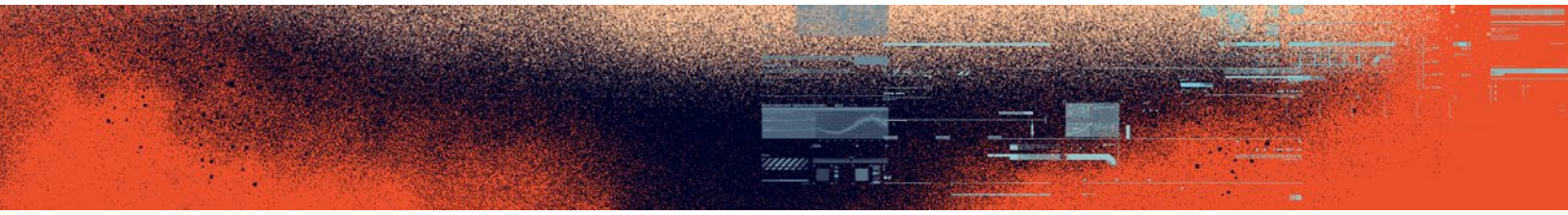
MI355X maintains a consistent throughput advantage across the full concurrency and TP sweep, with a median of 1.27x that reaches up to 2.54x at high tensor-parallel configurations.

Kimi-K2.6

This workload represents the single model tested where NVIDIA hardware demonstrated a consistent advantage for raw processing speed. At low-to-moderate request concurrency, the B200 outpaces the MI355X by a median of 17%.

However, this advantage is highly dependent on system load; at high concurrencies paired with extended context lengths, the AMD MI355X reverses the trend, outperforming the B200 by up to 1.96x. This pattern reflects MI355X's larger HBM capacity, which becomes a notable advantage at these demanding configurations, allowing it to maintain throughput while B200 faces memory pressure.

Signal65 Comment: *This report uses median values as representative figures, resulting in MI355X showing a throughput ratio of 0.86× for Kimi-K2.6. It should be noted, however, that when using a geometric mean the gap between MI355X and NVIDIA narrows to 0.96×. Under this measure, MI355X's price-performance advantage additionally expands to 1.59× more tokens per dollar, compared to 1.42× when computed using the median.*



Price-Performance

By applying the baseline GPU-hour infrastructure costs to the raw token processing rates, the raw processing power is converted into a standard economic unit of cost per 1 million tokens.

Cost per Million Tokens and Economic Advantage

The following table outlines the financial efficiency of both platforms across the evaluated large language models.

Model Architecture	AMD MI355X Cost / 1M Tokens (Median)	NVIDIA B200 Cost / 1M Tokens (Median)	AMD Tokens-per-Dollar Advantage (Median) [2]
GPT-OSS-120B	$0.091	$0.252	**2.15x more tokens/$**
Qwen3-Next-80B	$0.079	$0.188	**2.11x more tokens/$**
Kimi-K2.6	$1.18	$2.22	**1.42x more tokens/$**

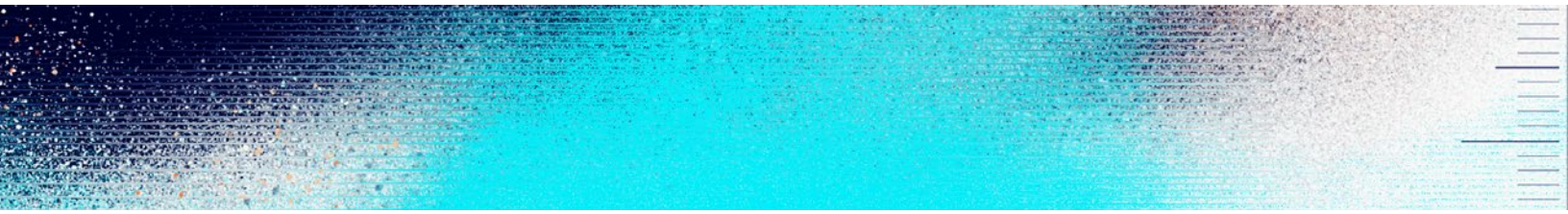
Figure 4: *Tokens-per-dollar Overview*

Signal65 Comment: *The cost-per-million-token figures above are derived from the blended on-demand GPU rental rates described in the Pricing Methodology section. They are not comparable to cost-per-token estimates based on the total cost of owning and operating hardware, which are calculated on a different basis.*

Key Economic Takeaways

Infrastructure costs fundamentally change the competitive picture for these systems. Because the blended on-demand rental rate for AMD MI355X capacity is approximately 40% lower than for B200, AMD achieves a clear price-performance victory across all three models when renting capacity from cloud providers.

The Kimi-K2.6 behavior highlights this shift. While the NVIDIA B200 generates tokens **17% faster** in raw hardware execution, the AMD MI355X produces 1.42x more tokens per dollar. For an enterprise operator purchasing compute to execute real workloads, the MI355X delivers a lower cost per unit of work across the board.



Real-World Business Outcomes

To ground these performance and economic metrics in practical operations, the data was mapped directly into three standard business workloads: two which involve batch document processing and one focused on concurrent interactive user serving. The workloads below are illustrative examples built on a set of stated, reasonable assumptions — they translate token throughput into business terms. Actual results will vary with workload characteristics, configurations, and requirements, though the underlying performance and cost patterns held across the configurations tested.

[2] Each column represents an independently calculated median across all tested configurations and concurrency levels. The cost ratio shown is the median of all cost ratios and cannot be directly computed from the median AMD and NVIDIA costs.

Batch Workload #1: Financial Analysis

This workload simulates an automated corporate portfolio analysis, evaluating user-specified companies using live financial data and technical indicators to generate structured investment reports. Each company analyzed in the report initiates its own inference call. Token profiles were measured empirically using a reference application.

Component	Description	Tokens
Business unit	100-company financial analysis	—
Input per company (ISL)	Full assembled prompt: template + stock summary + technical indicators + news headlines	~1,052
Output per company (OSL)	Structured analysis + recommendation	~720
Total per company	ISL + OSL	~1,772
Total per portfolio	100 companies × 1,772 tokens	~177,200

Figure 5: *Financial Analysis Workload Profile*

For this analysis, a single document represents a comprehensive batch assessment covering 100 distinct companies, consuming approximately 177,200 total tokens. The input and output token characteristics of each inference call closely match the 1024/1024 test shape used for GPT-OSS-120B testing. Because throughput rather than latency is the primary concern for batch workloads, higher concurrency levels are typical. A concurrency of 64 was chosen as the representative operating point, however other high concurrency levels tested show similar trends.

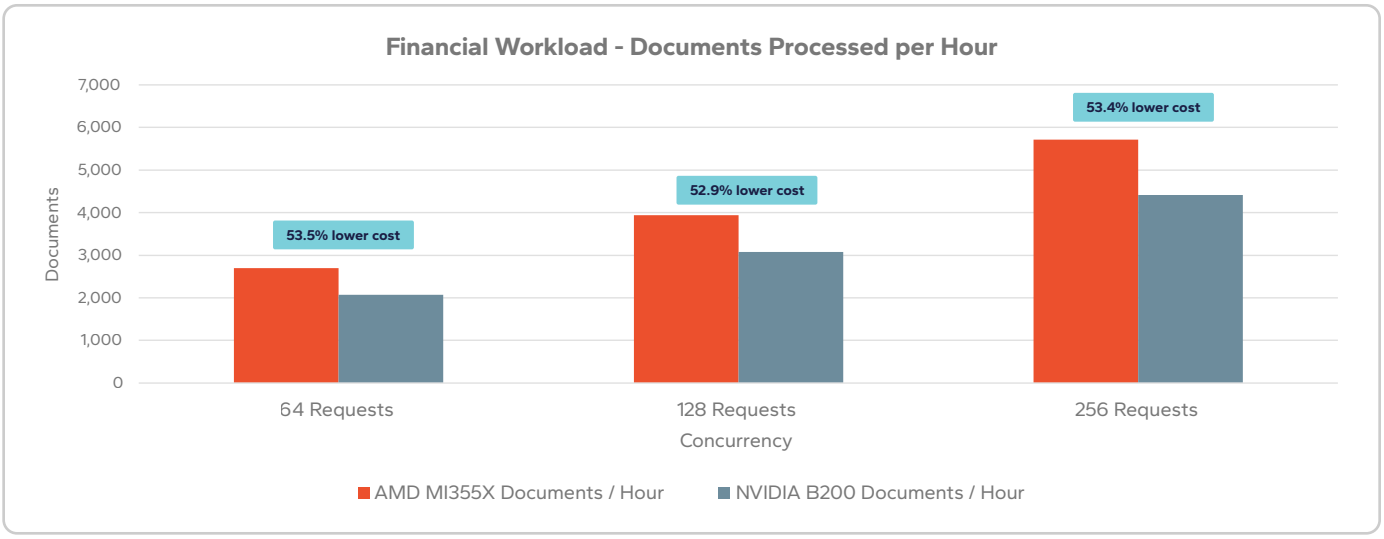


Figure 6: *Financial Analysis Workload Comparison*

Across all batch depths, the AMD MI355X cluster completes approximately 1.3x more documents per hour than the B200 deployment. When evaluating cost, the MI355X can run this financial analysis workload at less than half the cost per document of the NVIDIA infrastructure.

Batch Workload #2: Long Report Generation

A second batch workload, report generation, was analyzed to evaluate performance across a different token profile. For report generation, a large output is generated based on a relatively small input. This type of workload is closely represented by the 1024 / 8192 configuration tested on Kimi-K2.6. For this workload, a standard unit was defined as a single page of a document, with a page assumed to be approximately 405 tokens. Because this assumption applies identically to both platforms, changing the assumed tokens-per-page value rescales pages per hour and cost per page proportionally on both sides — the relative cost differences between platforms are unaffected.

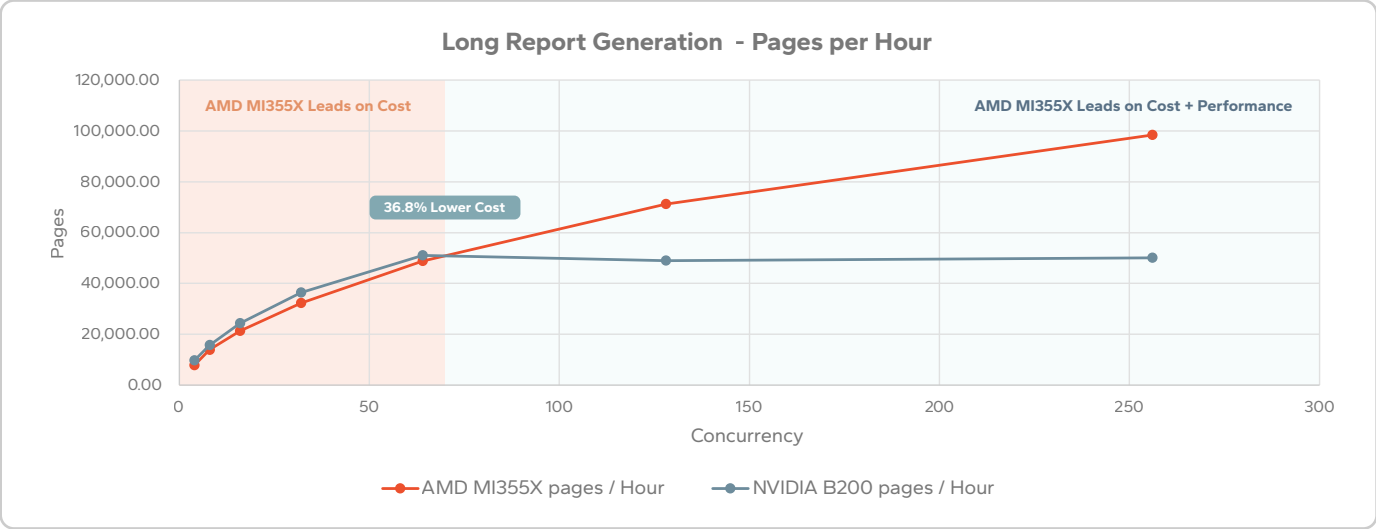


Figure 7: *Long Report Generation - Pages per Hour*

Similar to the first batch workload, a higher concurrency is preferable to maximize throughput, and a concurrency of 64 is again used as a representative operating point. This workload demonstrates how AMD MI355X can provide business outcome advantages, even without raw performance advantages. At a concurrency of 64, NVIDIA maintains a slight throughput advantage, of approximately 5%, enabling greater total report generation per hour. When considering the price to run the workload, however, AMD reclaims a significant advantage, costing 36.8% less per page.

At lower concurrency levels, NVIDIA B200 achieves greater throughput advantages, however, the MI355X maintains a strong price advantage, ranging from 25.11% to 31.96%. At the two highest concurrency levels, the MI355X's HBM advantage enables it to generate more pages than NVIDIA, resulting in further price advantages between 58.46% and 69.25%.

Interactive Serving Workload: RAG Chat Application

Interactive AI applications must maximize concurrent user capacity while keeping response latency below an acceptable threshold. Chatbots are a representative example — each query requires a quick response to maintain an acceptable user experience, making latency a first-order concern.

Token profiles were measured empirically using a reference RAG-enhanced chatbot application:

Component	Tokens
System prompt	~237
Retrieved RAG context (variable)	~530–1,250 (avg ~933)
Total input per turn (ISL)	~1,018 avg (range 673–1,515)
Output per turn (OSL)	~674 avg (range 462–876)

Figure 8: *RAG Chat Application Workload Profile (Single Turn)*

In production deployments, however, input lengths can grow much longer. Multi-turn conversation history and more extensive RAG retrieval can easily extend input sizes beyond what was measured in the reference application to 5,000 tokens or more. This analysis therefore uses Qwen3-Next-80B at the 5000/500 configuration as the representative workload shape. To approximate realistic latency requirements of interactive applications, this analysis imposed an end to end latency SLA of p99 ≤ 15 seconds.

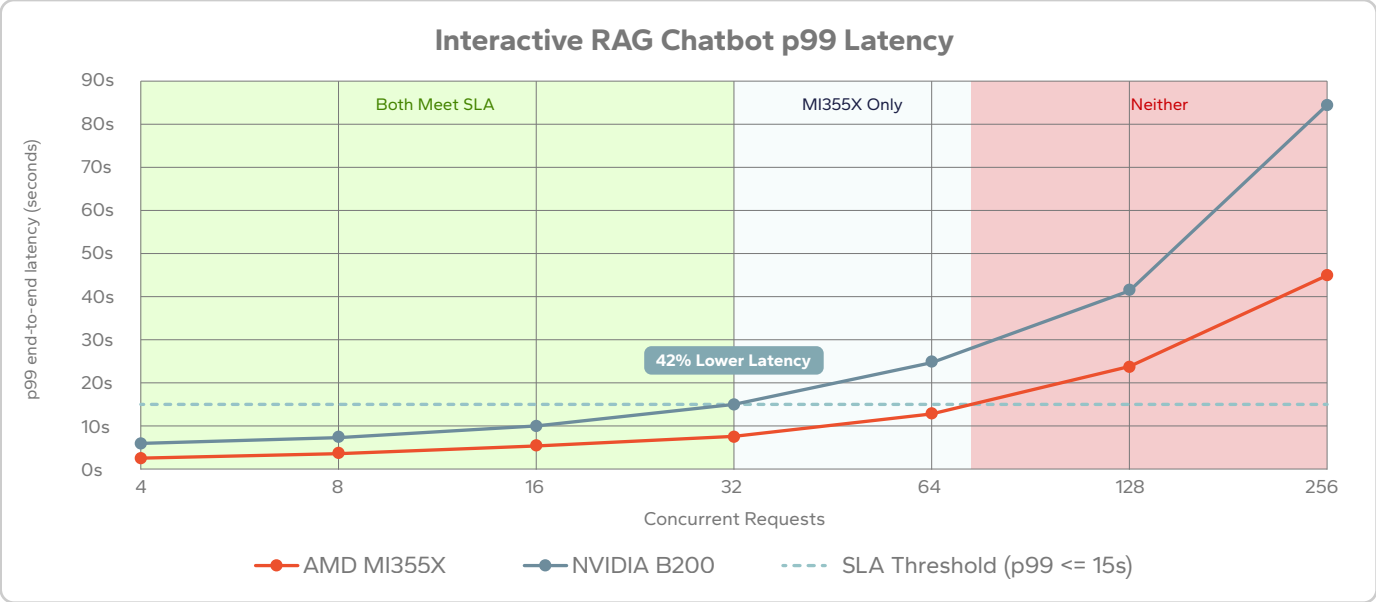


Figure 9: *Interactive RAG Chat Workload Comparison (5000 / 500 context size)*

At low concurrencies, both platforms meet the SLA and serve the same user population. At concurrency 32, B200 narrowly achieves the SLA by a margin of 85ms. At this same concurrency, MI355X serves the same number of users, but delivers 42% lower p99 latency and requires 39.6% lower cost per hour. At a concurrency of 64, MI355X remains within the SLA, while B200 exceeds the threshold. This indicates MI355X can sustain roughly 2x the concurrent users before violating the latency constraint. Given the same workload, this provides extra resilience for AMD MI355X to scale to double the user load, while NVIDIA B200 exceeds the SLA by 62% at 24.3 seconds.

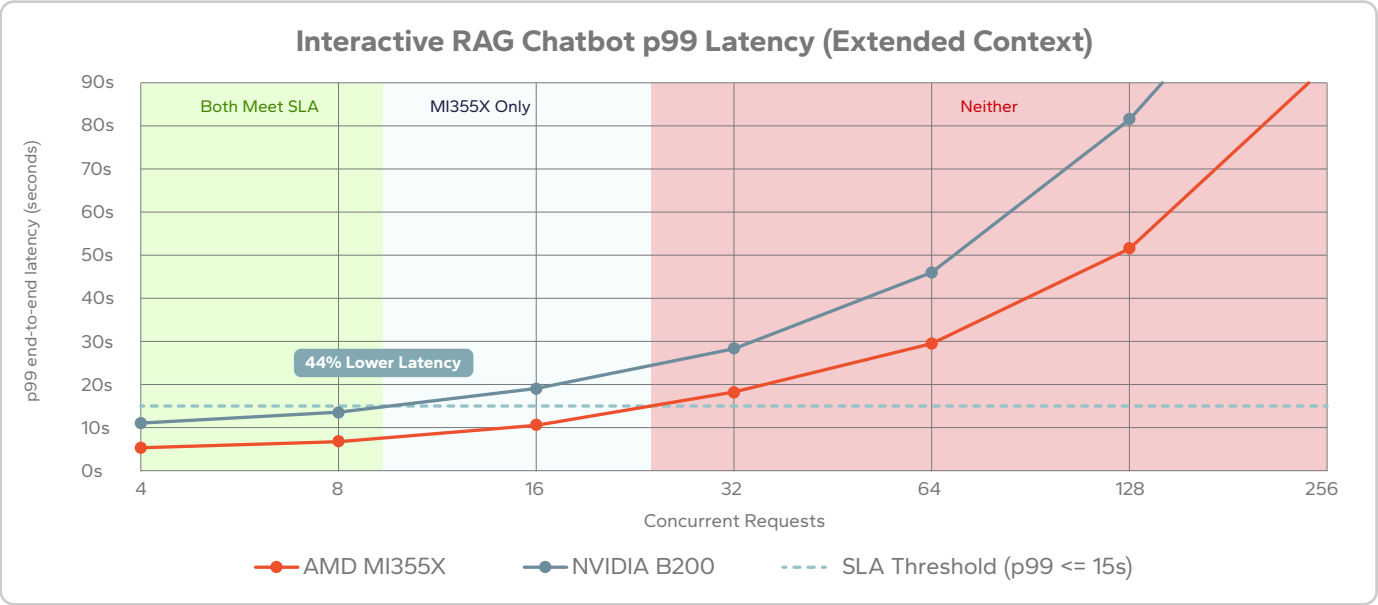


Figure 10: *Interactive RAG Chat Workload Comparison (10000 / 1000 context size)*

RAG workloads with longer input requirements are captured in the 10000/1000 test shape. At this larger context size, latencies are higher on both platforms, but the overall pattern is consistent with the 5000/500 results. At concurrency 8, both platforms meet the 15-second SLA, with MI355X delivering 44% lower p99 latency (7.9s vs 14.1s). At concurrency 16, MI355X remains within the SLA while B200 exceeds the threshold — consistent with MI355X's ability to sustain lower latency tail behavior as concurrency increases. The 39.6% cost-per-hour advantage carries through at all concurrency levels.

Signal65 Comment: *Concurrencies of 32 and 8 were chosen as representative configurations because they are B200's highest measured loads meeting the SLA threshold at each context size. The 5000/500 and 10000/1000 context size configurations were chosen as the best representatives of long-context interactive workloads. At the smaller 1000/100 context size configuration, AMD MI355X demonstrated even further concurrency under the SLA and larger advantages compared to NVIDIA B200.*

Strategic Business Conclusions

For enterprise buyers evaluating AI inference infrastructure, AMD MI355X presents a compelling case that extends well beyond benchmark headlines. The combination of competitive raw performance and significantly lower infrastructure cost means that organizations can do more for less money. This translates into processing more documents, serving more users, or generating more content, all at meaningfully lower cost. As AI workloads scale and infrastructure spending grows, the ability to deliver equivalent or superior outcomes at 39.6% lower cost per GPU-hour represents a decisive operational advantage.

Across all three evaluation dimensions included in this analysis, AMD MI355X delivers consistent advantages over NVIDIA B200:

- **Raw Performance:** MI355X leads on 2 of 3 models tested — 1.30x on GPT-OSS-120B and 1.27x on Qwen3-Next-80B, with NVIDIA B200 holding a ~17% throughput advantage on Kimi-K2.6
- **Infrastructure Cost:** MI355X carries a 39.6% lower GPU-hour rate ($4.71 vs $7.80), delivering 1.42x to 2.15x more tokens per dollar across all three models
- **Business Outcomes:** MI355X processes 1.3x more financial analysis documents per hour at 53% lower cost per document; for long-form report generation, it trades a ~5% throughput deficit for 37% lower cost per page; for interactive RAG serving, it delivers equivalent user capacity with ~42% lower latency at 39.6% lower cost per hour

Key Takeaways

- Raw performance leader on 2 of 3 models tested (1.30× and 1.27×)
- More tokens per dollar on all three models (1.42×–2.15×) — including Kimi-K2.6, where B200 is faster
- ~2× the concurrent users sustained within a 15s p99 SLA
- Infrastructure economics, not silicon speed alone, determine business value

As with any benchmark study, these findings reflect the specific conditions tested. On-demand GPU pricing moves with market supply and demand, and negotiated contract rates will differ from the blended spot rates used here — readers should apply their own pricing to the performance data when evaluating total cost. The results likewise reflect the three models and configurations tested; other workload shapes, models, and deployment choices will shift the specific ratios, even as the underlying patterns proved consistent across the configurations in this study. Finally, inference software for newly released models tends to mature over time — a dynamic worth watching on both platforms — and MI355X rental capacity is currently available from fewer cloud providers than B200, though that footprint is expanding. None of these considerations change the direction of the findings: they are the reasons Signal65 anchored every claim to disclosed configurations, published pricing, and reproducible data.

Critically, these advantages hold even on the model where NVIDIA leads on raw throughput — demonstrating that infrastructure economics, not silicon speed alone, determine real-world business value. For organizations making AI infrastructure decisions today, MI355X offers a rare combination of competitive performance and a cost structure that compounds into meaningful savings at scale.

Appendix A: Test Environment and Software Configuration

TABLE 1: Test Period and Hardware

Test Period	June 23 – July 10, 2026
AMD MI355X	8-GPU node hosted via TensorWave
NVIDIA B200	8-GPU node hosted via RunPod

All inference workloads were GPU-bound with models resident in GPU memory. Token throughput measurements are therefore isolated to GPU compute performance — host CPU, storage, and network are not in the measurement path. Any differences in host configuration between TensorWave and RunPod do not materially affect the reported throughput figures.

Request length variance: To simulate realistic workload conditions, request lengths were randomized around each target ISL/OSL value using a random range ratio of 0.8 (+/-20%) for GPT-OSS-120B and Kimi-K2.6, and 0.9 (+/-10%) for Qwen3-Next-80B.

TABLE 2: AMD MI355X Software Stack

Model	Inference Stack	Version	ROCm Version	PyTorch
GPT-OSS-120B	AMD ATOM	0.1.2.post	7.2.2	2.10.0
Qwen3-Next-80B	vLLM-ROCm nightly	0.23.1rc1 (commit: 04c2a8de)	7.2.3	—
Kimi-K2.6	vLLM-ROCm nightly	0.23.1rc1 (commit: 04c2a8de)	7.2.3	—

TABLE 3: NVIDIA B200 Software Stack

Model	Inference Stack	Version	CUDA Architecture
GPT-OSS-120B	TensorRT-LLM	1.2.1	10.0 (Blackwell)
Qwen3-Next-80B	vLLM	0.22.1	10.0 (Blackwell)
Kimi-K2.6	vLLM	0.21.0	10.0 (Blackwell)

TABLE 4: Model Precision

Model	AMD MI355X Precision	NVIDIA B200 Precision
GPT-OSS-120B	MXFP4 weights + FP8 KV cache	MXFP4 weights + FP8 KV cache
Qwen3-Next-80B	FP8 weights + FP8 KV cache	FP8 weights + FP8 KV cache
Kimi-K2.6	MXFP4 weights, FP8 KV cache	NVFP4 weights, FP8 KV cache

Appendix B: Methodology Notes & Scope Boundaries

Load Generation Discrepancy Note

The standard benchmarking toolkit used was vllm bench serve. However, the NVIDIA TensorRT-LLM stack on gpt-oss-120b required a modified bench_openai.py script. Production validation showed this script yields a **5% to 17% higher throughput** reading than the standard vllm bench serve engine under identical conditions. This variance favors the NVIDIA B200 cluster, meaning the reported AMD performance advantage on gpt-oss is inherently conservative.

Explicit Data Inclusions and Exclusions

Cold Start Outlier Exclusion: A single specific data point—Kimi-K2.6, 1024 input / 1024 output tokens at a request concurrency of 4—was omitted from the summary metrics. During this initial run, the B200 cluster encountered an un-warmed cold start, resulting in an anomalous throughput of 88.5 tokens/sec/GPU compared to over 410 tokens/sec/GPU in adjacent testing blocks. Removing this clear artifact is a conservative adjustment; keeping the anomalous data point would skew the median results to make the NVIDIA B200 look ~3% worse overall on the Kimi model. This specific point is preserved in the raw appendix data tables for audit transparency, marked explicitly as (excl).

Appendix C: Business Workloads - Detailed Results

Financial Analysis - Documents per Hour

Concurrency Level	AMD MI355X Documents / Hour	NVIDIA B200 Documents / Hour	AMD Cost per Document	NVIDIA Cost per Document	Cost Difference
64 Requests	2,693	2,075	$0.0140	$0.0301	53.46%
128 Requests	3,942	3,075	$0.0096	$0.0203	52.89%
256 Requests	5,715	4,411	$0.0066	$0.0141	53.40%

Long Report Generation - Pages per Hour

Concurrency	AMD MI355X pages / Hour	NVIDIA B200 pages / Hour	AMD MI355X $/page	NVIDIA B200 $/page	Cost Difference
4	7,786.67	9,656.89	0.004839	0.006462	25.11%
8	13,916.44	15,680.00	0.002708	0.003980	31.96%
16	21,297.78	24,334.22	0.001769	0.002564	31.01%
32	32,206.22	36,480.00	0.001170	0.001711	31.60%
64	48,760.89	51,036.44	0.000773	0.001223	36.80%
128	71,196.44	48,974.22	0.000529	0.001274	58.46%
256	98,410.67	50,112.00	0.000383	0.001245	69.25%

Interactive RAG Chat Workload Comparison (5000 / 500 context size)

Concurrency	MI355X p99	B200 p99	SLA ≤ 15s
4	3.0s	6.0s	Both
8	4.0s	7.2s	Both
16	5.6s	10.1s	Both
32	8.6s	14.9s	Both
64	13.8s	24.3s	MI355X only
128	24.6s	42.0s	Neither
256	44.9s	84.0s	Neither

Interactive RAG Chat Workload Comparison (10000 / 1000 context size)

Concurrency	MI355X p99	B200 p99	SLA ≤ 15s
4	6.0s	11.7s	Both
8	7.9s	14.1s	Both
16	11.5s	19.5s	MI355X only
32	18.0s	28.4s	Neither
64	29.5s	46.4s	Neither
128	53.0s	81.8s	Neither
256	96.3s	163.5s	Neither

Appendix D: Detailed Configuration Throughput Tables

GPT-OSS-120B Performance Results

TP	ISL	OSL	CONC	MI355X ATOM tok/s/ GPU	B200 TRT tok/s/ GPU	Ratio (MI355X/ B200)	MI ATOM p99 e2e (ms)	B200 TRT p99 e2e (ms)
TP=1	1024	1024	4	1963.1	1745.1	**1.125**	5827	5292
TP=1	1024	1024	8	3670.1	3347.7	**1.096**	6749	5238
TP=1	1024	1024	16	6386.6	5643.4	**1.132**	7791	5993
TP=1	1024	1024	32	10918.9	8455.8	**1.291**	8711	7946
TP=1	1024	1024	64	16567.7	12767.9	**1.298**	10956	10539
TP=1	1024	1024	128	24254.3	18920.7	**1.282**	16846	14074
TP=1	1024	1024	256	35166.1	27137	**1.296**	19152	20622
TP=1	1024	8192	4	1146.4	959.6	**1.195**	36928	39166
TP=1	1024	8192	8	2233.9	1908	**1.171**	38026	39286
TP=1	1024	8192	16	3974.1	3182.5	**1.249**	42807	47777
TP=1	1024	8192	32	6431.4	4642	**1.385**	53090	66001
TP=1	1024	8192	64	10834.5	6695.9	**1.618**	63232	90554
TP=1	1024	8192	128	14350.5	9537.6	**1.505**	95573	125332
TP=1	1024	8192	256	19297.8	8587.6	**2.247**	149675	294451
TP=1	8192	1024	4	8175.9	7095.8	**1.152**	5741	5334
TP=1	8192	1024	8	14917.4	10385.7	**1.436**	7140	7329

TP=1	8192	1024	16	23683.9	17505.4	**1.353**	9475	8913
TP=1	8192	1024	32	33597.5	25742.6	**1.305**	12921	13425
TP=1	8192	1024	64	48093.9	33461.8	**1.437**	18943	22298
TP=1	8192	1024	128	55612.5	40311.5	**1.38**	32995	44225
TP=1	8192	1024	256	62249.3	40667.7	**1.531**	59339	73058

Qwen3-Next-80B Results

TP	ISL	OSL	CONC	MI355X vLLM tok/s/ GPU	B200 vLLM tok/s/ GPU	Ratio (MI355X/ B200)	MI355X vLLM p99 e2e (ms)	B200 vLLM p99 e2e (ms)
TP=1	1000	100	4	7014.6	3963.1	**1.77**	632	2636
TP=1	1000	100	8	11396.7	7115.7	**1.602**	777	2890
TP=1	1000	100	16	17628.5	11502.5	**1.533**	1069	3059
TP=1	1000	100	32	26257.8	15629.3	**1.68**	1614	4345
TP=1	1000	100	64	35379.8	22334.3	**1.584**	2608	6130
TP=1	1000	100	128	43633.2	26773.7	**1.63**	4749	10129
TP=1	1000	100	256	51403	23881	**2.152**	8426	24862
TP=1	5000	500	4	7467.2	6235.1	**1.198**	2992	6011
TP=1	5000	500	8	11642.7	10621.5	**1.096**	4009	7205
TP=1	5000	500	16	17803.8	15992.8	**1.113**	5644	10097
TP=1	5000	500	32	26134	21803.2	**1.199**	8598	14915
TP=1	5000	500	64	34259.6	28371.9	**1.208**	13806	24280
TP=1	5000	500	128	41393.9	32532	**1.272**	24609	42027
TP=1	5000	500	256	48881.5	34180	**1.43**	44898	84003
TP=1	10000	1000	4	7391	6373.7	**1.16**	6049	11703

TP=1	10000	1000	8	11544.1	10915.7	**1.058**	7884	14084
TP=1	10000	1000	16	17718.4	16639.2	**1.065**	11493	19453
TP=1	10000	1000	32	24782	22901.5	**1.082**	18031	28396
TP=1	10000	1000	64	32310.5	29902.1	**1.081**	29476	46439
TP=1	10000	1000	128	39220.4	33834.8	**1.159**	53002	81759
TP=1	10000	1000	256	45252.6	35470.2	**1.276**	96281	163457
TP=2	1000	100	4	3791.9	2090.7	**1.814**	586	1784
TP=2	1000	100	8	6954.4	3234.8	**2.15**	652	3456
TP=2	1000	100	16	10740.6	5528	**1.943**	880	3195
TP=2	1000	100	32	16289.2	7712.6	**2.112**	1295	5133
TP=2	1000	100	64	23695.4	13341.1	**1.776**	1807	5182
TP=2	1000	100	128	30754.2	16060.2	**1.915**	3225	8887
TP=2	1000	100	256	36323.3	20699.9	**1.755**	5955	15335
TP=2	5000	500	4	4170.3	3168.4	**1.316**	2652	5681
TP=2	5000	500	8	7211.7	5558.6	**1.297**	3187	7133
TP=2	5000	500	16	11017.7	9169.7	**1.202**	4449	8557
TP=2	5000	500	32	17125.3	13418.8	**1.276**	6415	11890
TP=2	5000	500	64	23535.2	18434.8	**1.277**	10198	19212
TP=2	5000	500	128	29663.8	23669.1	**1.253**	17501	28923
TP=2	5000	500	256	34327.7	27808.4	**1.234**	32144	51955
TP=2	10000	1000	4	4123.1	3394.5	**1.215**	5371	10704
TP=2	10000	1000	8	7175.7	5992.9	**1.197**	6390	12641
TP=2	10000	1000	16	10802.3	9527.6	**1.134**	9330	17354
TP=2	10000	1000	32	16479.3	13963.5	**1.18**	13247	23502

TP=2	10000	1000	64	22473.4	20176.5	**1.114**	21431	33513
TP=2	10000	1000	128	28057.2	24537.9	**1.143**	37228	56775
TP=2	10000	1000	256	32241.8	27410.8	**1.176**	68415	109177
TP=4	1000	100	4	1913.3	1105.2	**1.731**	584	1742
TP=4	1000	100	8	3675.7	1925.1	**1.909**	610	2060
TP=4	1000	100	16	6350.4	2829	**2.245**	752	3126
TP=4	1000	100	32	10368.2	4083.8	**2.539**	982	5394
TP=4	1000	100	64	14261.4	6544	**2.179**	1473	6500
TP=4	1000	100	128	20264.5	9176.4	**2.208**	2483	7375
TP=4	1000	100	256	24398.2	12959.6	**1.883**	4449	11343
TP=4	5000	500	4	2131.5	1752.5	**1.216**	2594	5164
TP=4	5000	500	8	3847.3	3197.4	**1.203**	2935	5815
TP=4	5000	500	16	6572.9	5159.3	**1.274**	3711	7652
TP=4	5000	500	32	10483.3	7136	**1.469**	5088	12000
TP=4	5000	500	64	14817	11371.9	**1.303**	8011	14790
TP=4	5000	500	128	19324.7	14600.5	**1.324**	13507	26603
TP=4	5000	500	256	22842	18266.6	**1.25**	24303	39891
TP=4	10000	1000	4	2112.2	1817.8	**1.162**	5238	10047
TP=4	10000	1000	8	3799.5	3423.3	**1.11**	5949	10958
TP=4	10000	1000	16	6430.9	5621.1	**1.144**	7702	13853
TP=4	10000	1000	32	10103.7	8470.5	**1.193**	10817	18605
TP=4	10000	1000	64	14107.3	11722.8	**1.203**	17037	31098
TP=4	10000	1000	128	17939	15142.4	**1.185**	29026	45117
TP=4	10000	1000	256	21063	18223.4	**1.156**	52412	79535

Kimi-K2.6 Performance Results

TP	ISL	OSL	CONC	MI355X vLLM tok/s/ GPU	B200 vLLM tok/s/ GPU	Ratio (MI355X/ B200)	MI355X vLLM p99 e2e (ms)	B200 vLLM p99 e2e (ms)
TP=4	1024	1024	4	188.7	88.5	**2.133 (excl)**	17250	150480
TP=4	1024	1024	8	343.3	411	**0.835**	19371	16452
TP=4	1024	1024	16	509.1	653.5	**0.779**	28709	21222
TP=4	1024	1024	32	821.6	1019.9	**0.806**	36087	27171
TP=4	1024	1024	64	1240.8	1534.8	**0.808**	46517	37196
TP=4	1024	1024	128	1717.9	2208.7	**0.778**	67034	51050
TP=4	1024	1024	256	2351.9	2884.5	**0.815**	100398	79360
TP=4	1024	8192	4	109.5	135.8	**0.806**	143713	116002
TP=4	1024	8192	8	195.7	220.5	**0.888**	154821	140092
TP=4	1024	8192	16	299.5	342.2	**0.875**	204084	182081
TP=4	1024	8192	32	452.9	513	**0.883**	274728	244557
TP=4	1024	8192	64	685.7	717.7	**0.955**	367529	359302
TP=4	1024	8192	128	1001.2	688.7	**1.454**	502743	650347
TP=4	1024	8192	256	1383.9	704.7	**1.964**	735953	1030685
TP=4	8192	1024	4	714.4	937.3	**0.762**	19472	16180
TP=4	8192	1024	8	1295.2	1547.7	**0.837**	26208	19894
TP=4	8192	1024	16	1918.3	2289.6	**0.838**	39384	27889
TP=4	8192	1024	32	2757.7	3101.3	**0.889**	52914	43210
TP=4	8192	1024	64	3778.3	3518	**1.074**	77998	67729
TP=4	8192	1024	128	4740.9	3566.7	**1.329**	126468	108995
TP=4	8192	1024	256	5750.7	3611.9	**1.592**	217375	192747

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