signal65

BREAKING THE MILLION-TOKEN BARRIER:

The Business Impact of Azure ND GB300 v6 Performance for Enterprise AI

AUTHOR

Russ Fellows
Principal Analyst | Signal65

NOVEMBER 2025

IN PARTNERSHIP WITH

Microsoft

Executive Summary: Enterprise AI at Scale

Enterprises are entering a new era of generative AI where success depends less on experimentation and more on execution at scale. The ability to move from pilot deployments to production-ready systems hinges on whether infrastructure can deliver both the raw performance and the sustained efficiency required for complex, always-on AI workloads. Until recently, the kind of throughput necessary to power thousands of concurrent users, large retrieval-augmented generation (RAG) pipelines, or multi-step agentic systems was available only to hyperscalers and research institutions.

In testing validated by Signal65, Microsoft Azure has demonstrated **an aggregate LLM inference throughput of 1,100,948 tokens per second** on a single rack of its next-generation ND GB300 v6 virtual machine infrastructure, powered by 72 NVIDIA Blackwell Ultra GPUs. This milestone is significant not just for breaking the one-million-token-per-second barrier and being an industry-first, but for doing so on a platform architected to meet the dynamic use and data governance needs of modern enterprises. Azure provides the foundational capabilities, from data residency controls and sovereign landing zones to robust encryption and confidential computing, that allow organizations to deploy powerful AI workloads while ensuring their data remains within their specified geographical and security boundaries.

This achievement fundamentally alters the calculus of AI efficiency and returns by proving that performance and compliance are not mutually exclusive. The key findings of this analysis are:

- **Unprecedented Application Scale Within a Compliant Framework:** The demonstrated throughput can support thousands of concurrent user interactions per second on a platform designed to meet complex regulatory requirements, enabling the deployment of at-scale AI inference services in sensitive industries.
- **Superior Generational Efficiency:** The Azure ND GB300-based platform delivers a 27% inference performance uplift over the previous NVIDIA GB200 generation for only a 17% increase in its power specification. Compared to the NVIDIA H100 generation, NVIDIA GB300 NVL72 offers nearly a 10x increase for inference performance at a nearly 2.5x power efficiency gain when measured at rack level. This yields significant improvements in performance-per-watt, which translate directly to a lower Total Cost of Ownership (TCO) and a more sustainable footprint for secure AI workloads.
- **No-Compromise CSP:** No other major cloud provider has published any MLPerf-like Llama 2 70B inference results near this scale. The latest v5.1 submissions achieved roughly 100,000 tokens per second on an 8-GPU DGX B200 configuration, nearly 10x slower than Azure's validated rack-scale result.
- **Enterprise-Grade Resilience and Stability:** The milestone was achieved over a sustained 80-minute benchmark run, proving the platform's stability for mission-critical, 24/7 production environments where reliability and data integrity are paramount.
- **The Democratization of Secure AI Supercomputing:** This achievement signifies that elite AI performance is no longer the exclusive domain of hyperscale AI companies. It is now an accessible, on-demand utility for mainstream enterprises through the Azure cloud, lowering the barrier to entry for building the next generation of AI applications.

This report will deconstruct this performance milestone, providing a detailed technical analysis of the system and methodology before translating these findings into their tangible business implications. The conclusion is clear: performance at this level is a key enabler for the next wave of AI innovation, particularly for the complex, agentic systems.

Validating Rack-Scale Inference on Microsoft Azure

To substantiate a claim of this magnitude, a rigorous and transparent testing methodology is essential. This section details the system, the workload, and the results of the benchmark, which was observed and validated by Signal65 acting as an independent third party. This test is not an official submission to MLCommons, but the test was executed in strict adherence to the MLPerf framework to ensure the results are credible, industry-relevant, and reproducible.

The System Under Test (SUT): Architecture for Resilient Scale

The test was conducted on a full NVL72 rack of Azure ND GB300 v6. This is not a monolithic supercomputer but is architected as **18 distinct NDv6 virtual machine (VM) instances** running in parallel. Each of these VMs is a powerful unit, equipped with four NVIDIA Blackwell Ultra GPUs, two NVIDIA Grace™ CPUs, local NVMe storage, and the necessary networking (NVIDIA NVLink™ inside a rack and NVIDIA InfiniBand between racks) to operate as part of the larger cluster, bringing the total to 72 GPUs.

This architectural choice is a critical differentiator that reflects a cloud-native approach to high-performance computing. Traditional HPC systems often rely on large, monolithic nodes where the failure of a single component, such as a GPU or network interface, can render the entire multi-GPU node unusable. The Azure architecture, by contrast, is designed for resilience. According to Microsoft's engineering team, the system is intentionally divided into smaller, four-GPU fault domains. If one VM instance experiences a hardware failure, the remaining instances and their 68 GPUs continue to operate unaffected. This design minimizes the "blast radius" of a potential failure, ensuring a higher degree of service continuity, which is a non-negotiable requirement for enterprise-grade production systems. Therefore, the 1.1 million token-per-second result is significant not just for its raw performance but for the resilient and fault-tolerant way it was achieved.

The Benchmark: Real-World Enterprise Demand

The credibility of any performance claim rests on the relevance of the workload used for testing. The configuration for this benchmark was carefully selected to represent the demanding generative AI applications that enterprises are actively developing and deploying.

- **Workload:** The test utilized the **Llama2-70B model**, a powerful 70-billion-parameter open-source LLM that has become an industry-standard benchmark for evaluating systems on complex reasoning and generation tasks, and the OpenOrca data set. Its size and capability ensure that the performance results are representative of high-value, real-world use cases.
- **Precision and Optimization:** The model was run using NVFP4 precision, a form of quantization that significantly accelerates inference speed while maintaining high accuracy. This was implemented via **NVIDIA's TensorRT-LLM** library, a highly optimized, production-ready software stack for LLM inference. The use of standard, "out-of-the-box" software optimizations is crucial, as it demonstrates a level of performance that Azure customers can realistically achieve without bespoke, non-transferable engineering efforts.
- **Testing Harness:** The **MLCommons MLPerf Inference v5.1 toolset** provided the testing framework [User Query]. The test was run in the **Offline Scenario**, which is specifically designed to measure the maximum peak experimental throughput of a system. In this scenario, the entire dataset of prompts is sent to the System Under Test in a single batch, effectively saturating the hardware to reveal its peak processing capacity. This makes it the ideal metric for understanding the absolute performance ceiling of the infrastructure and for capacity planning in high-volume environments.

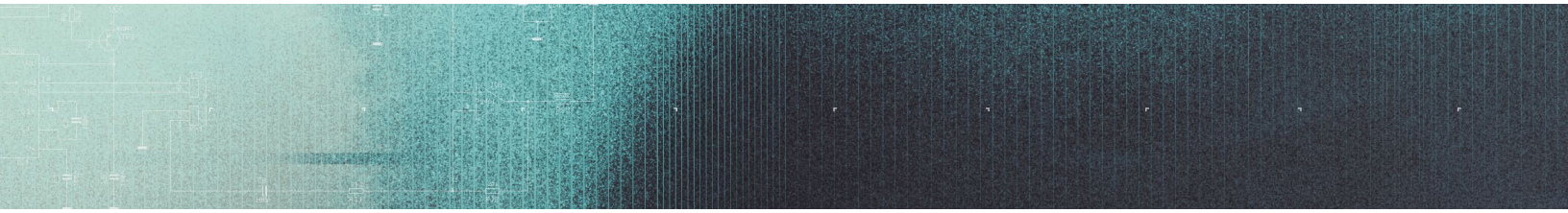
The Result: 1.1 Million Tokens per Second

Across the 18 parallel ND GB300 v6 VM instances, the system achieved a total aggregate throughput of **1,100,948.3 tokens per second**. This figure is the sum of the individual performance results from each of the 18 four-GPU nodes. The detailed performance breakdown, shown in Table 1, provides transparency and context to this aggregate number, substantiating the claim that this is a true rack-scale result rather than a theoretical extrapolation from a single test. This empirical, multi-node validation stands in contrast to methodologies that simply multiply the result of a single best-case run, offering a more realistic and credible depiction of at-scale performance.

Table 1: Rack-Scale Performance Summary (Llama2-70B, NVFP4)

Metric	Performance (Tokens/Second)
Total Aggregate Throughput	1,100,948.3
Maximum Single-Node Throughput	62,804
Minimum Single-Node Throughput	57,599
Average Single-Node Throughput	61,164
Median Single-Node Throughput	61,759

Source: *Data compiled from 18 parallel test runs observed by Signal65.*



Performance Analysis: Stability and Efficiency

A headline performance number, while impressive, tells only part of the story. For enterprise decision-makers, the consistency, stability, and efficiency of that performance are equally, if not more, important. This section provides a deeper analysis of the results to assess the production-readiness of the Azure platform.

Sustained Throughput: The Importance of Resilience

The MLPerf Offline benchmark is not a short sprint; the Signal65-validated test ran for approximately **1 hour and 20 minutes at more than 1 million samples per query**. Achieving stable, high throughput results over a long duration is a critical data point. It serves as a rigorous stress test, demonstrating that the performance is not a fleeting, unsustainable peak but a consistent and reliable output. For enterprise applications that must operate continuously, this proves the system's thermal stability, the robustness of the software stack, and the platform's ability to handle prolonged, maximum-intensity load without performance degradation or failure. This sustained performance is a key indicator of the infrastructure's readiness for mission-critical, 24/7 inferencing scenarios.

Understanding Performance Variation: Engineering vs. Production

The results in Table 1 show a performance delta of approximately 8.9% between the fastest-performing node (62,804 tokens/sec) and the slowest-performing node (57,599 tokens/sec). While any variation warrants examination, this result must be understood within its proper context. The test was conducted on a "pre-production" environment used for engineering, testing, and validation. Such clusters inherently have more performance variability due to concurrent activities including software updates, debugging, and other tests running in the background. But with the median and average values as close in proximity as they are, this suggests general throughput stability with minimal spikes or throughs in performance.

Viewed through this lens, achieving this level of performance and a relatively modest variation on a pre-production system is a strong positive signal. Historical data from more mature Azure offerings provides a reliable forecast for what customers can expect from generally available (GA) instances. For example, similar tests on production Azure clusters running the previous-generation NVIDIA Blackwell GPUs showed a much tighter performance variation of just 1.7% to 3%. It is therefore reasonable to project that as the GB300-based offering moves into full production and is deployed in optimized, isolated customer environments, the performance will be even more consistent and potentially higher than what was demonstrated here. This addresses the variability not as a weakness, but as a testament to the high performance achieved even on a non-production system, reinforcing the value of Azure's rigorous engineering and quality control processes that precede a public launch.

Business Translation: From Tokens to Enterprise Value

So does 1.1 million tokens per second create concrete business value? High-throughput inference is the key to unlocking cost-effective AI at scale, directly impacting user experience, application capacity, and the financial return on AI investments.

Modeling Concurrent User Capacity

For a business leader, the most pressing question is: "What does this performance mean for my application and my users?" By modeling common enterprise AI workloads, it is possible to translate the aggregate throughput into an estimate of concurrent user capacity. These models are based on typical token counts observed in real-world applications, from simple chatbots to more complex RAG and agentic systems.

Table 2 provides estimates for the number of simultaneous user interactions a full NVL72 rack of Azure ND GB300 v6 VMs could process per second for four distinct workload profiles.

Table 2: Estimated Concurrent User Capacity at 1.1M Tokens/Second

Workload Profile	Description	Estimated Tokens per Interaction	Estimated Concurrent Interactions per Second
Simple Q&A	A single user question and a direct response, typical of a customer service chatbot.	650 (150 in, 500 out)	~1,690
RAG Query	User query plus retrieval of 3 context chunks (384 tokens each) and a synthesized answer.	2,500 (150 in, 1150 ctx, 1200 out)	~440
Document Summarization	User submits a 2,000-token document (e.g., a report) and receives a 500-token summary.	2,500 (2,000 in 500 out)	~440
Multi-Step Agentic Task	A complex task requiring 4 conversational turns of reasoning and tool use to complete a workflow.	10,000	~110

Note: *Token assumptions are derived from analysis of various AI and RAG applications through Signal65 engagements with third-parties and enterprise customers. "Concurrent Interactions per Second" is calculated as 1,100,000 divided by "Assumed Tokens per Interaction." Results are batched and don't include interactivity assumptions.*

This analysis makes the performance tangible. It allows a CIO or a line-of-business leader to see that a single Azure rack could potentially serve nearly 1,700 simultaneous chatbot users per second or support over 400 concurrent, complex RAG queries every second. This level of capacity from a single compute unit fundamentally changes the economics of deploying high-demand AI services to a large enterprise workforce or a global customer base. In practice, these levels of concurrency could support significantly more active users throughout a typical work day.

Generational Leaps in Efficiency: Performance-per-Watt

Cost-effective scaling is driven not just by raw performance, but by the efficiency of that performance. For large-scale AI deployments, power consumption is a significant and growing component of operational expenditure. Analyzing the performance-per-watt reveals a crucial advantage of the new Azure platform.

The test data shows that the GB300-based system delivers a **27% performance increase** over the previous-generation GB200 system (1.1M vs. 865k tokens/sec). This substantial performance gain is achieved with only a **17% increase in GPU power specification** (1400W for GB300 vs. 1200W for GB200). The result is a marked improvement in efficiency, or the number of tokens processed for every watt of energy consumed. Compared to the widely deployed H100 generation, the efficiency gains are even more pronounced, with the new platform delivering approximately double the tokens-per-watt.

This superior efficiency is a strategic advantage. For enterprises, this translates into:

1. **Lower Operating Costs:** Reduced energy consumption leads to lower electricity bills, directly impacting the TCO of the AI infrastructure.
2. **Increased Compute Density:** More computational work can be done within the same power and thermal envelope of a data center, maximizing the value of the physical footprint.
3. **Sustainable AI:** A more energy-efficient platform supports corporate sustainability goals by reducing the environmental impact of large-scale AI operations.

Table 3 quantifies this generational leap in efficiency, providing a metric for evaluating the ROI of migrating workloads to the latest infrastructure.

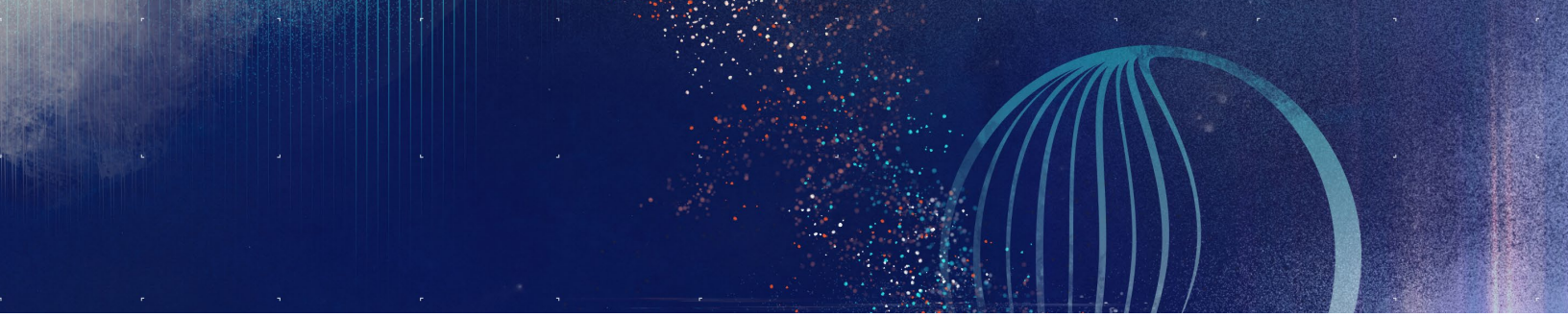
Table 3: Generational Efficiency Comparison (Inference Performance-per-Watt, rack level)

GPU Generation	Rack Performance (Tokens/Sec)	Rack Power Draw (kW)	Efficiency (Tokens/Watt)	Generational Uplift vs. H100
NVIDIA H100 (32 GPUs)	98,000	22.4	~4.39	Baseline
NVIDIA Blackwell (72 GPUs)	865,000	86.4	~10.01	+128%
NVIDIA Blackwell Ultra (72 GPUs)	**1,100,948**	**100.8**	**~10.92**	**+148%**

Note: *Rack power draw calculated as (Number of GPUs) x (TDP per GPU). H100 performance is an estimate based on public data for comparable workloads. GB200 and GB300 data from test validation.*

Unlocking Advanced AI Workloads

The combination of massive throughput and improved efficiency makes previously niche or cost-prohibitive AI workloads viable for mainstream enterprise deployment. The most significant of these are complex, multi-step agentic systems. These AI agents, which can reason, plan, and execute tasks by interacting with other systems and tools, consume significantly more tokens than simple Q&A bots. The performance demonstrated here provides the necessary capacity to run thousands of such agents concurrently, paving the way for a new generation of applications in areas like process automation, sophisticated data analysis, and autonomous customer service workflows. This infrastructure milestone is the foundational enabler for the strategic vision of an "explosion" of AI use driven by agents.



Conclusion: The Democratization of Supercomputing-Class AI

The achievement of over 1.1 million tokens per second on a single Azure rack is more than a benchmark record; it is a definitive proof point that the performance required for large-scale, transformative AI is achievable as a reliable, efficient, and resilient utility. This milestone **effectively democratizes access to supercomputing-class AI**, moving it from the exclusive domain of specialized research labs and hyperscale AI developers into the hands of mainstream enterprises.

Through a potent combination of NVIDIA's cutting-edge GB300 NVL72 system, a resilient cloud-native architecture that prioritizes fault tolerance, and a highly optimized software stack with TensorRT-LLM, Microsoft Azure is systematically dismantling the barriers to production AI. The analysis shows that this platform not only **delivers unprecedented scale** but does so with superior energy efficiency, directly addressing the critical TCO and sustainability concerns of modern enterprises. The sustained, stable performance over a prolonged test period provides the confidence needed to build mission-critical applications on this foundation.

Currently, Signal65 is not aware of any other cloud vendor delivering a system that can approach this level of Llama2-70B inferencing performance. Azure stands alone in providing a resilient, and flexible AI supercomputing class system accessible to IT users in a production cloud environment.

This performance benchmark should not be viewed as an endpoint, but rather as a new starting line. It provides the capacity for enterprises to move beyond exploratory AI projects and begin engineering the next generation of intelligent applications. From powering company-wide knowledge management systems with RAG to deploying fleets of AI agents to automate complex business processes, this level of performance enables businesses to innovate with confidence, secure in the knowledge that the underlying platform can meet the demands of their most ambitious AI initiatives.

Appendix: Test Configuration and Disclosures

System Under Test (SUT) Configuration

Component	Specification
Cloud Platform	Microsoft Azure
VM Instance SKU	ND_GB300_v6 (Canary Cluster)
System Configuration	18 x NDv6 VM instances in a single NV72 rack
GPU	4 x NVIDIA GB300 per VM (72 total)
GPU Memory	189,471 MiB per GPU
GPU Power Limit	1,400 Watts
Storage	14 TB Local NVMe RAID per VM
Software Stack	Azure CycleCloud
LLM Inference Engine	NVIDIA TensorRT-LLM [1]
Benchmark Harness	MLCommons MLPerf Inference v5.1 [User Query]
Benchmark Scenario	Offline
Model	Llama2-70B
Precision	NVFP4

MLPerf Disclosure Statement

Unverified MLPerf® v5.1 Inference Closed Llama 2-70B offline. Result not verified by MLCommons Association. Unverified results have not been through an MLPerf review and may use measurement methodologies and/or workload implementations that are inconsistent with the MLPerf specification for verified results. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use is strictly prohibited. See www.mlcommons.org for more information. Results obtained using NVIDIA MLPerf v5.1 code with TensorRT-LLM 0.18.0.dev.

A Note on Latency: Throughput vs. Responsiveness

The detailed logs from the MLPerf run report latency metrics on the order of many minutes, with a median end-to-end latency of approximately 38 minutes (2,328 seconds). It is critical to understand that these figures are **not representative of user-facing latency** or Time-to-First Token (TTFT). They are an artifact of the MLPerf Offline scenario's specific methodology.

The Offline test is structured to process a large batch of samples (specifically 1,056,000) as a single, continuous query that runs for the entire duration of the benchmark (approximately 80 minutes). The reported "latency" is the time taken to complete this massive, monolithic query, not the time an end-user would wait for a response to a single prompt. The MLCommons framework includes entirely separate benchmark scenarios, namely the Server and Interactive scenarios, which are specifically designed to measure user-centric latency metrics like TTFT and Time-Per-Output-Token (TPOT) under various concurrency constraints. This clarification is essential to prevent misinterpretation of the results and demonstrates a sophisticated understanding of the benchmarking process: the Offline test correctly measures maximum throughput, which was the goal of this exercise.

Link to Public Results and Reproduction Guide

For full transparency and to allow for independent review, the 18 individual result log files from this test, along with a guide on the methodology used, are publicly available in the Azure AI Benchmarking Guide GitHub repository at the following location:

https://github.com/Azure/AI-benchmarking-guide/tree/main/Azure_Results/1M_ND_GB300_v6_Inference

Works Cited

1. What runs ChatGPT, Grok, DeepSeek, Llama & agents? - YouTube, accessed October 22, 2025, https://www.youtube.com/watch?v=Jy7hopkFIYg
2. Llama 2 70B: An MLPerf Inference Benchmark for Large Language Models - MLCommons, accessed October 22, 2025, https://mlcommons.org/2024/03/mlperf-llama2-70b/
3. AI trends 2025: Adoption barriers and updated predictions - Deloitte, accessed October 22, 2025, https://www.deloitte.com/us/en/services/consulting/blogs/ai-adoption-challenges-ai-trends.html
4. Data, privacy, and security for Azure Direct Models in Azure AI ..., accessed October 22, 2025, https://learn.microsoft.com/en-us/azure/ai-foundry/responsible-ai/openai/data-privacy
5. AI workloads and sovereignty - Microsoft Sovereign Cloud ..., accessed October 22, 2025, https://learn.microsoft.com/en-us/industry/sovereign-cloud/sovereign-public-cloud/implementing-workloads/ai-workloads-sovereignty
6. Microsoft Cloud for Sovereignty: a solution for companies concerned ..., accessed October 22, 2025, https://www.devoteam.com/expert-view/microsoft-cloud-for-sovereignty-a-solution-for-companies-concerned-about-their-sovereignty/

Important Information About this Report

CONTRIBUTORS
Russ Fellows
Principal Analyst | 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.

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

Microsoft

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