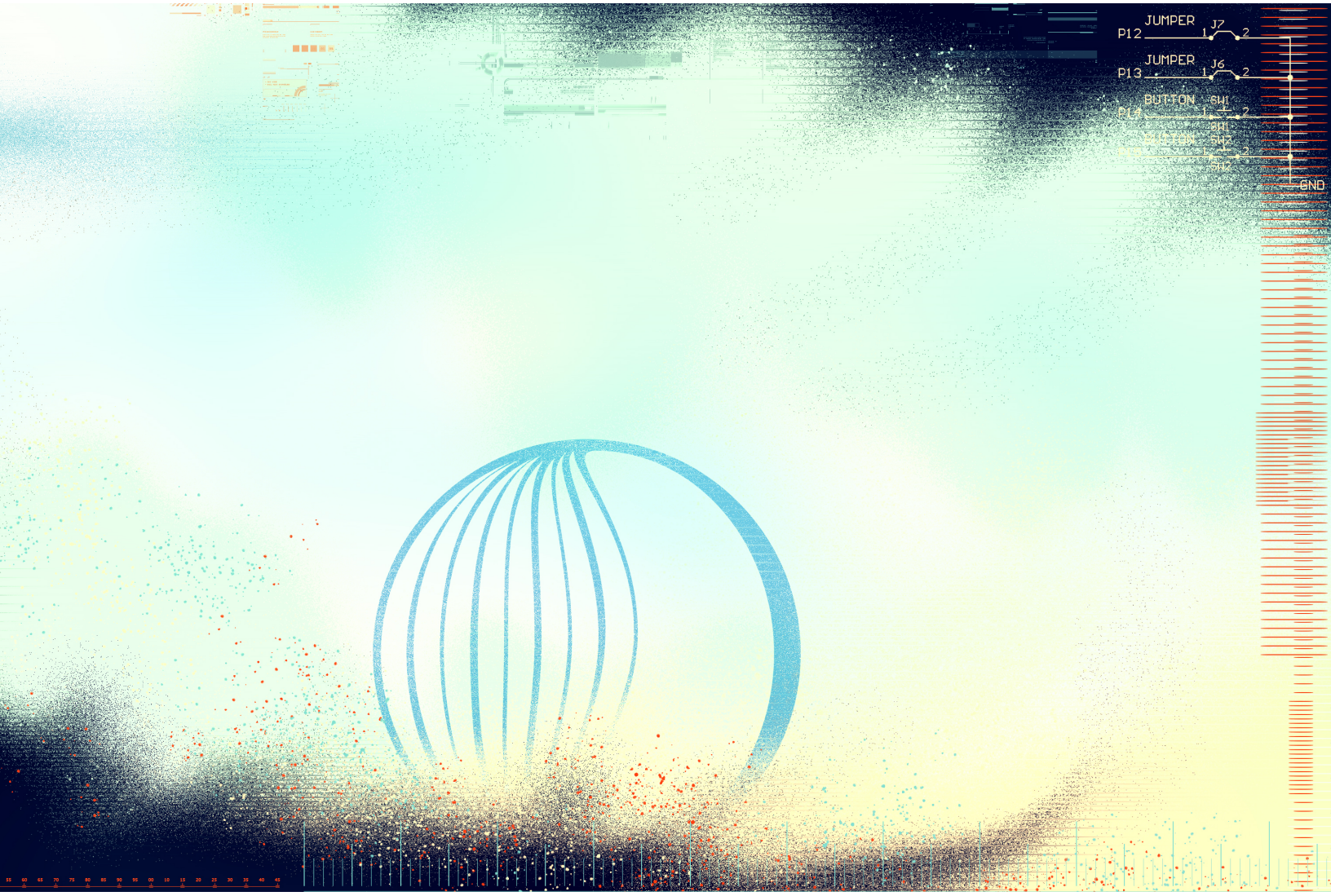
signal65



VERIFIED LEADERSHIP:

Analyzing Microsoft Azure ND GB200 v6 VMs Inference Performance

AUTHOR

Russ Fellows
VP - Labs | Signal65

IN PARTNERSHIP WITH

Microsoft

SEPTEMBER 2025

Executive Summary

As the focus of generative AI for enterprises moves experimental deployments to production inferencing deployments, efficiency is a critical aspect to achieving business value. In this environment, infrastructure decisions require rigorous, transparent and comparable performance data for comparison.

Against that backdrop, the latest MLPerf Inference v5.1 results provide a vendor neutral view of how platforms perform under standardized workloads. In this paper, we focus on the Microsoft Azure ND GB200 v6 virtual machines (VMs) accelerated by the NVIDIA GB200 NVL4. We analyze Azure's MLPerf results in the Llama2 70B and Llama3.1 405B benchmarks, explain why these tests matter for real deployments, and translate raw data into actionable insights.

Microsoft Azure is a leading provider of AI infrastructure in the public cloud. The Azure ND GB200 v6 is a generally available accelerated instance designed with a scale-out NVIDIA NVLink fabric and platform software stack that targets both training and inference.

This report analyzes the latest MLPerf Inference v5.1 along with relevant v5.0 results, focusing on the performance of Microsoft Azure's ND GB200 v6 virtual machines, which are accelerated by the NVIDIA GB200 NVL72. The analysis reveals a clear and verifiable leadership position for the Azure platform.

Key findings include:

- **Audited Leadership on a Critical Workload:** On the industry-standard Llama-2-70B benchmark, the Azure ND GB200 v6 virtual machine demonstrates leadership in tokens-per-second per GPU within the NVIDIA Blackwell accelerator class. The platform delivered audited results of 13,015.40 tokens/s per accelerator in the Offline scenario.
- **Lower Total Cost of Ownership (TCO):** Through continual hardware and networking stack optimizations, along with a constantly tuned software stack to provide predictable scaling and reduced operational risk for mission-critical AI deployments.
- **Demonstrated Platform Optimization:** Azure has achieved a significant 8.3% performance improvement on the Llama-2-70B Offline benchmark compared to previously published figures. This gain underscores a mature and continuous optimization cadence across the entire platform stack, delivering increasing value to customers on the same hardware.
- **Unmatched Transparency in the Cloud Market:** Microsoft Azure is one of only a few major cloud vendors to submit results for the NVIDIA Blackwell accelerator class, providing a high degree of transparency for comparing cloud hosted GPU options.
- **Validated Readiness for Frontier Models:** By submitting audited results for the Llama-3.1 405B model, Azure validates its platform's capability to handle the next generation of frontier-scale AI models, offering customers crucial investment protection for their future roadmaps, with 211.81 tokens/s per accelerator for Offline scenario.

Ultimately, these benchmark results are more than technical metrics; they are direct indicators of business value. As outlined, Azure's verified performance translates into tangible TCO benefits, which in turn provides Azure clients with one of the most efficient AI stacks available.

Delivering AI Value: The Inference Imperative

Today, enterprises across industry segments are focused on the operational challenge of deploying these AI models at scale to create tangible business value. From powering internal copilots that enhance employee productivity to driving customer-facing autonomous agents in contact centers, the act of serving trained models, or inferencing has become the critical path to realizing the return on AI investments.

This transition to inferencing efficiency, while providing high quality results via large and frontier class models has changed the decision-making metrics that define success. The key performance indicators that matter in production are the sustained throughput of tokens-per-second delivered at a defined quality of service, combined with the ability to scale throughput predictably to ensure enterprise AI applications remain resilient, and economically viable.

MLPerf Datacenter Benchmarks

In this new landscape, theoretical specifications like datasheet Tera Operations Per Second (TOPS) are insufficient. Enterprises require a clear, objective understanding of how a complete platform, from the silicon, to networking, storage and the software stack all perform under realistic workloads. This need for verifiable, performance data is a challenge that the MLPerf benchmarks help address, by providing a fair basis for evaluating solutions.

This report focuses exclusively on results from the **Closed Division**, which mandates that all submitters use the exact same reference model and pre-trained weights. This constraint is critical because it isolates the performance of the underlying platform—the hardware and software stack—making this a good choice for fair, and unbiased comparisons.

For enterprises deploying large language models (LLMs), the MLPerf Inference Datacenter suite provides two particularly relevant scenarios that simulate distinct but equally important production environments: the Offline scenario and the Server scenario.

Decoding MLPerf Datacenter Scenarios

The **Offline scenario** is designed to measure the maximum raw throughput of a system. In this test, the entire batch of inference requests is sent to the system under test (SUT) at once. There is no strict latency requirement for individual requests; the objective is to process the full batch as quickly as possible. This scenario is an excellent proxy for non-interactive, batch-processing workloads. Enterprise use cases include the large-scale analysis of document archives, offline summarization of research papers, or the pre-generation of personalized content where immediate responsiveness is not the primary concern. The key metric is pure throughput, typically measured in tokens per second.

In contrast, the **Server scenario** simulates an interactive online service. Inference requests arrive not as a single batch but at random intervals that follow a Poisson distribution, mimicking the arrival of user queries to a web

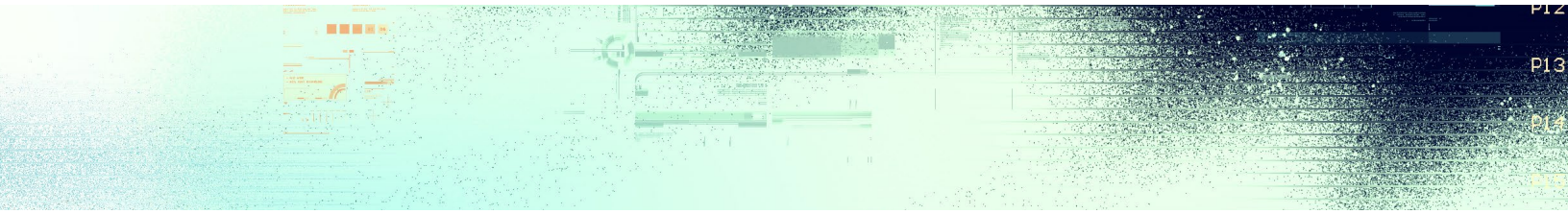
service. Each request must be completed within a strict latency target to ensure a responsive user experience. This scenario is important for user-facing generative AI applications, such as interactive chatbots, real-time translation APIs, or enterprise copilots embedded in productivity software. The metric of success is not just raw speed, but the maximum rate of queries (or tokens per second) the system can sustain while ensuring that 99% of requests meet the latency service level agreement (SLA).

A system's performance in the Server scenario is typically lower than in the Offline scenario because the latency constraint prevents perfect, large-scale batching.

Table 1 provides a summary of these two critical datacenter scenarios and their relevance to enterprise AI deployments.

Scenario	Workload Simulation	Key Constraint	Primary Metric	Enterprise Use Case Examples
Offline	All queries are sent to the system at once in a single batch.	None - Maximize Throughput	Tokens per Second	Batch document analysis, offline report generation, large-scale data classification.
Server	Queries arrive in a random, unpredictable stream (Poisson distribution).	Strict Latency Target (e.g., 99th percentile latency).	Queries/Tokens per Second at Target Latency	Interactive chatbots, real-time API services, enterprise copilots, live translation.

Table 1: *MLPerf Inference Datacenter Scenarios Explained*



Analysis of MLPerf v5.1 Results: Azure Performance on the NVIDIA GB200 Platform

This analysis focuses on the official results from the MLPerf Inference v5.1 benchmark round, providing a quantitative assessment of Microsoft Azure performance. The evaluation centers on the Azure ND GB200 v6 virtual machine, as detailed in submission ID 5.1-0008, and maintains a strict focus on the NVIDIA GB200 NVL72 and closely related NVIDIA HGX B200 accelerator class to ensure a true, like-for-like comparison of next-generation AI infrastructure.1

The data in this analysis are the official MLPerf Inference v5.0 and 5.1 results. To facilitate fair comparisons between systems of varying sizes, for example a 4-GPU virtual machine versus a larger, multi-node

configuration, all performance metrics have been normalized to a **per-GPU tokens/second** basis. This normalization provides a clear and direct measure of the underlying efficiency of the platform's architecture and software stack, independent of the total number of accelerators deployed. By isolating per-accelerator efficiency, enterprises can more accurately project performance and cost at any scale.

Azure Llama-2-70B Leadership: Performance and Progress

The Llama-2-70B model has become the de facto industry standard for open-weights LLMs and, as of the latest MLPerf rounds, has surpassed traditional vision models like ResNet-50 as the most submitted benchmark.1 Its widespread adoption for enterprise fine-tuning and deployment makes it the most critical benchmark for evaluating the performance of mainstream generative AI infrastructure.

In the MLPerf v5.1 submission, the Azure ND GB200 v6 VM delivered exceptional performance on this key workload. The platform achieved a throughput of **52,061.6 tokens/s** in the Offline scenario. When normalized to a per GPU basis, this translates to **13,015.4 tokens/s/GPU (Offline)**. The results vs. other Cloud competitors are shown below in Figure 1.

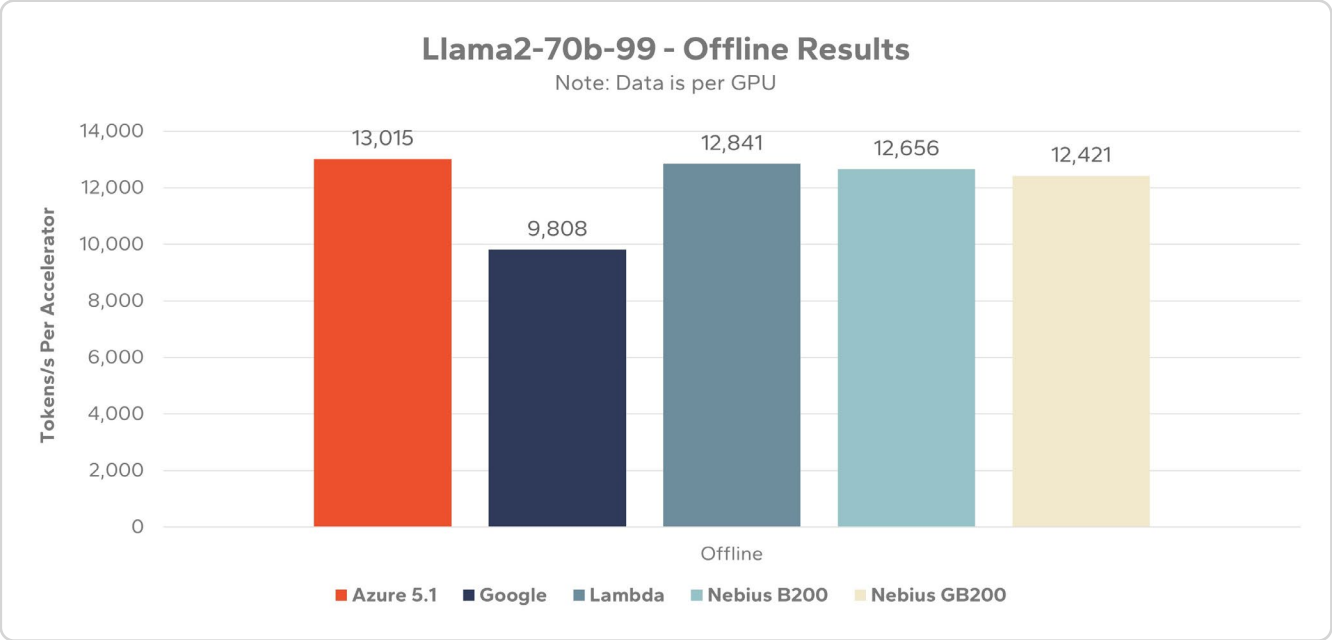


Figure 1: *Llama2-70b-99 MLPerf version 5.0 and 5.1 Results (Source: MLCommons Org)*

Validating for the Future: Llama-3.1-405B Results

While Llama-2-70B represents the current mainstream, the Llama-3.1-405B benchmark serves as a crucial stress test for the future. With 405 billion parameters and support for vastly larger context windows, this model pushes the boundaries of memory bandwidth and interconnect performance, representing the class of frontier models that enterprises will seek to deploy next. Additionally, Llama-3.1-405B is a different type of model, known as a Mixture of Expert (MoE) model. These typically require larger memory to load the entire model, but then activate only a portion of their nodes, according to the chosen experts identified for the task

being inferenced. This allows comparatively faster inferencing, due to requiring only a subset of the model weights to be utilized at any one time.

By submitting a valid, audited Offline result of 847.23 tokens/s, or 211.81 tokens/s per VM, Azure is demonstrating more than just performance; it is demonstrating this platform is ready for very large frontier models. This provides customers with a powerful form of investment protection, validating that the infrastructure they procure today is being actively tested and optimized for the models they will need tomorrow.

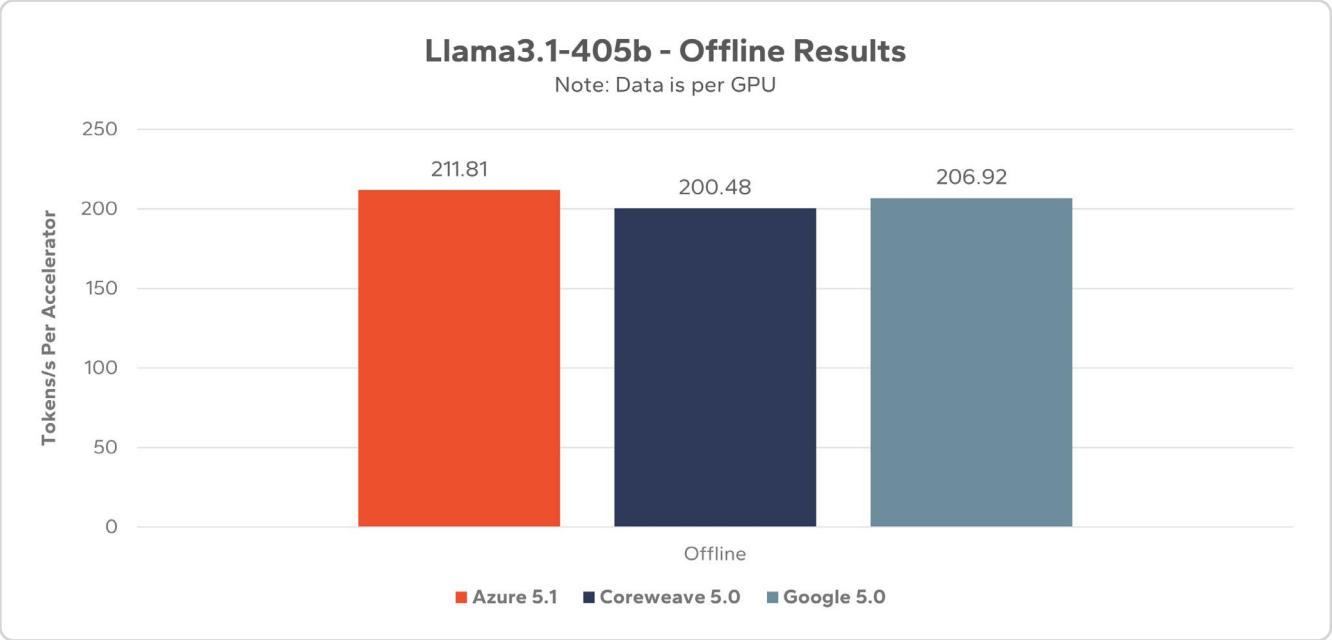


Figure 2: *Llama3.1-405b MLPerf version 5.0 and 5.1 Results (Source: MLCommons Org)*

The Cloud Competitive Landscape: The Power of Transparency

The most compelling aspect of Azure's MLPerf v5.1 submission lies not only in the numbers themselves but in the context of the competitive landscape. Azure's own engineering blogs detail the architectural choices (e.g., GB200 NVL72 scale units) and the single-node validation methodology used to ensure that improvements discovered on one VM translate to predictable gains across racks and datacenters.

Based on a thorough analysis of the final, official results file for the GB200/B200 accelerator subset, Microsoft Azure is one of only two major cloud providers with published and audited results.

This fact alters the nature of procurement and vendor evaluation. In the absence of comparable, publicly vetted data from other cloud providers on the same class of hardware, any competing performance claims remain unverified. An enterprise making a significant infrastructure investment decision based on a vendor's private, internal testing is shouldering a significant degree of performance risk.

Azure's decision to submit results provides transparency, and effectively shifts the burden of proof to competitors, who must justify the absence of equivalent public data. In this context, Azure's audited MLPerf submission becomes a strategic asset that directly translates to lower risk and greater confidence for its customers.

Translating Performance to Enterprise Value

The technical metrics and competitive positioning detailed in the MLPerf benchmarks are not academic exercises; they have direct and material consequences for the business outcomes of enterprise AI initiatives. The leadership performance demonstrated by Azure's ND GB200 v6 platform translates into tangible advantages in cost, operational stability, and strategic planning.

The most direct financial benefit of superior inference performance is a reduction in TCO. The metric of per-GPU efficiency (tokens/s/GPU) is a clear indicator of how much workload a single unit of hardware can handle. A higher efficiency means an organization can serve the same number of users or process the same volume of data with a smaller fleet of accelerators. This consolidation directly reduces capital expenditures on hardware, as well as the associated operational costs of power, cooling, and data center space.

For example, the 8.3% performance uplift Azure demonstrated on the Llama-2-70B benchmark is a material gain at scale. For an organization planning a deployment that would have previously required 12 virtual machines, that same workload can now be handled by just 11. Across a large-scale deployment, this seemingly small percentage gain compounds into substantial savings, allowing organizations to either reduce their budget or reinvest the savings into expanding their AI capabilities. At rack scale, these benefits continue to accrue, with a full NVIDIA GB200 NVL72 rack of 18 GB200 nodes providing approximately 1.5x more processing for additional workloads.

Production AI services are expected to be available and performant 24/7. Performance degradation or outages can have immediate impacts on customer experience and revenue. Problems like thermal throttling, where performance degrades as hardware heats up, or software stack bottlenecks that emerge only under continuous load, are often surfaced by the rigorous testing process. By successfully completing these audited runs, Azure demonstrates a level of operational resilience that gives enterprises confidence in the platform's stability. Azure provides recipes and guidance for consistently achieving high performance on their AI infrastructure nodes.[1]

One of the greatest challenges in scaling AI services is accurately predicting resource needs to handle fluctuating user demand. Over-provisioning leads to wasted expenditure on idle resources, while under-provisioning results in poor performance and a negative user experience. Azure's engineering practice of linking single-node validation, as seen in the MLPerf VM results, to sustained telemetry from full-rack deployments provides a crucial bridge. It gives customers confidence that the performance observed on a single VM will scale predictably across an entire cluster. This allows for more precise and reliable capacity planning, enabling organizations to build right-sized environments that can confidently absorb traffic spikes without being wastefully over-engineered.

[1] Link to: Azure AI Performance & Benchmarking Guide

Conclusion and Recommendations

The analysis of the MLPerf Inference v5.1 results clearly demonstrates that the Microsoft Azure ND GB200 v6 platform provides a high performing and scalable foundation for enterprise-grade generative AI inferencing. Azure's ND results in the MLPerf Inference v5.1 (ID 5.1-0008) demonstrate measurable platform progress and strong like-for-like performance within the GB200/B200 accelerator class. The Llama-2-70B Offline result improved by ~8.3% versus the prior public figure released 6 months ago of 865,584 tokens/s. The new results of ~937,109 tokens/s per NVIDIA GB200 NVL72 rack. Per-GPU efficiency similarly improved to 13,015.4 tokens/s/GPU. For enterprises deploying generative AI applications, these gains translate directly into less hardware for the same workload. Reducing hardware translates directly to lower costs both for on premises and cloud hosted workloads.

Azure's performance is distinguished by several key factors. First, it demonstrates leadership efficiency on the Llama-2-70B model, the most relevant benchmark for today's mainstream enterprise workloads. Second, the platform shows a consistent cadence of improvement, with an 8.3% performance gain that signals mature, ongoing optimization delivering compounding value to customers. Third, by successfully submitting results for the frontier-scale Llama-3.1-405B model, Azure validates its readiness for the next generation of AI, offering customers vital investment protection. Finally, and most critically, Azure's commitment to transparency—as the one of the few major cloud providers with audited GB200/B200 results in the v5.1 benchmark round provides performance assurance.[2]

Based on these findings, the following strategic recommendation is offered to enterprises:

Organizations should request potential partners to provide verifiable results on the specific accelerator class being considered for deployment. Performance claims based on internal, unverified benchmarks, or on different hardware classes, should be treated as introducing significant risk. By establishing this high standard for evidence, enterprises can:

- **Dramatically reduce performance risk** by ensuring that infrastructure choices are based on data proven to be realistic and reproducible.
- **Improve negotiating leverage** by requiring all vendors to compete on a level playing field of verifiable, apples-to-apples data.
- **Make informed, data-driven decisions** that align infrastructure investments with tangible business outcomes like lower TCO and greater operational resilience.

Microsoft Azure has met this rigorous standard, providing a clear, data-backed foundation for building the next generation of enterprise AI applications.

[2] Azure Blog on High Performance GB200 Results

Important Information About this Report

CONTRIBUTORS

Russ Fellows
VP, Labs | 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

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