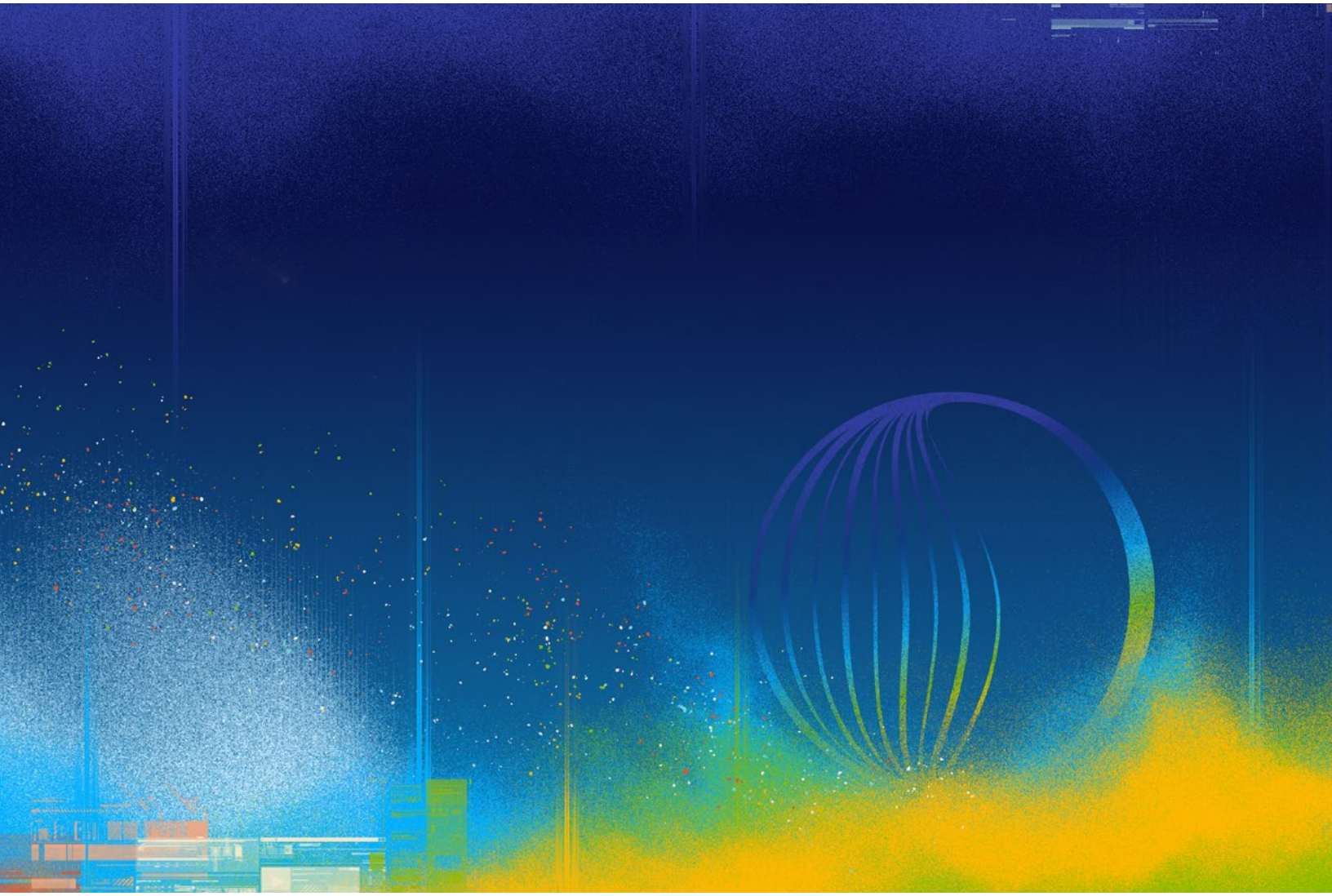
signal65



Microsoft Azure Leads DeepSeek-R1 Interactive Inference in MLPerf v6.1

AUTHORS

Russ Fellows
VP, Labs | Signal65

IN PARTNERSHIP WITH

Microsoft

NOVEMBER 2026

Executive Summary

Independent Signal65 analysis of audited DeepSeek-R1 results shows Azure's ND GB300 v6 submission, built on NVIDIA GB300 NVL72, leading Interactive throughput by 10.3% over the next cloud submission.

Key Highlights

- **10.3% Interactive lead** over the next cloud submission (260,098 output tokens per second)
- **93.6% and 96.4% scaling efficiency** from one rack to four (Server and Offline)
- **2.09 million output tokens per second** in Offline at 288 GPUs
- **38.1% Server uplift** from GB200 to GB300 at 72 GPUs
- Azure is the only cloud provider with a multi-rack NVL72 DeepSeek-R1 result in this round

The MLCommons MLPerf Inference v6.1 results provide an important reference point for organizations evaluating inference options for large AI systems. As enterprises move from standard dense models to reasoning-centric models based on mixture-of-experts (MoE) architectures and longer generation cycles, datacenter infrastructure must support greater compute, memory and networking demands. MLPerf Inference v6.1 offers an audited, objective basis for comparing these capabilities under consistent operating scenarios. Signal65, a leading independent third-party AI benchmarking and validation firm, analyzed the audited published results for Microsoft Azure's DeepSeek-R1 submissions and compared them with other Available systems at matched GPU generation and count. This is Azure's first DeepSeek-R1 submission to MLPerf, and it includes a measured four-rack, 288-GPU ND GB200 v6 system, built on NVIDIA GB200 NVL72, alongside single-rack ND GB200 v6 and ND GB300 v6 results. Azure is the only cloud provider in this round with a multi-rack NVL72 result for DeepSeek-R1, which gives customers a direct measurement of multi-rack capacity.

Microsoft Azure's MLPerf Inference v6.1 results support Azure as an inference platform for both single-rack deployments and larger installations, combining Interactive leadership, competitive single-rack scale-up throughput and nearly linear scale-out across four racks. Together, these results give infrastructure teams two measured Azure capacity points, one for a competitive single-rack platform and one for a four-rack, 288-GPU deployment. The clearest scale-out result comes from Azure ND GB200 v6, where expanding from 72 to 288 GPUs delivers 3.75x the Server throughput and 3.86x the Offline throughput, equal to 93.6% and 96.4% of ideal fourfold scaling efficiency, respectively.

Azure also leads the MLPerf results for DeepSeek-R1 Interactive throughput. Its 72-GPU ND GB300 v6 submission reaches 260,098 output tokens per second, which is 10.3% above the next-highest cloud submission and 2.6% above NVIDIA's own GB300 reference. These comparisons use the same model, scenario, GPU generation and GPU count. MLPerf measures three scenarios. Interactive applies the tightest latency limits and is most relevant to applications where users wait on generated output, such as conversational tools and agentic coding assistants. Server measures throughput as requests arrive over time, and Offline measures batch throughput.

© 2026 Signal65. All rights reserved.

Four implications stand out for organizations evaluating inference platforms:

- **Near-linear scaling beyond one rack.** The measured 288-GPU GB200 system produces 1.60 million tokens per second in Server and 2.09 million in Offline. Azure retains most of its single-rack throughput per GPU as the tested system grows from one to four racks.
- **Leadership in strict interactive tests.** Azure's GB300 Interactive score is the highest among the five Available Interactive submissions analyzed. Interactive applies substantially tighter response-time limits than the Server category.
- **Strong results at same GB200 scale.** Azure leads all four Available 72-GPU GB200 systems in both Server and Offline. Its margins over the closest cloud submission, from a large neocloud provider, are 1.7% and 2.3%, respectively.
- **Higher throughput with GB300.** At 72 GPUs, Azure GB300 increases Server throughput by 38.1% and Offline throughput by 22.0% over Azure GB200 in the same benchmark round.

These findings are specific to the MLPerf scenarios analyzed in this report, each of which uses different measurement conditions and, for Server and Interactive, different latency service level objectives (SLOs). Azure's 72-GPU GB300 Server and Offline results sit within 2.3% and 4.2% of the highest scores in a four-system peer group whose results span less than 5%. All results in this report come from MLPerf's Available category, which excludes preview and research systems.

All throughput figures are aggregate output tokens per second. Selected result IDs, calculations and five public references appear in the appendix.

Understanding the Results

MLPerf provides a common workload and test procedure for comparing complete inference systems. This report analyzes Azure's DeepSeek-R1 submissions and comparable systems in the MLPerf Inference v6.1 results. The benchmark evaluates a large reasoning model whose generated responses place sustained demands on the inference platform.

Three Measured Scenarios

Scenario	What it measures	DeepSeek-R1 latency SLOs (99th percentile)
Interactive	Throughput under tighter response constraints	TTFT 1.5 s TPOT 15 ms
Server	Throughput with requests arriving over time	TTFT 2.0 s TPOT 80 ms
Offline	Throughput for a batch of requests	No latency SLO

Time to first token (TTFT) measures how long a request waits for its first generated token. Time per output token (TPOT) measures the average interval between subsequent generated tokens for a response. The latency-constrained scenarios apply 99th-percentile requirements. A throughput score represents the system's aggregate output under those test conditions, rather than an individual user's generation speed.

© 2026 Signal65. All rights reserved.

Interactive applies the tightest latency SLOs of the three scenarios. It measures how much aggregate inference work a system can support under those limits, which makes it the most relevant result for applications where users wait for generated output, including conversational tools and coding assistants.

Interactive, Server and Offline should be compared separately. DeepSeek-R1 Interactive permits the specified speculative-decoding optimization, which predicts and verifies candidate tokens. Its score should not be treated as the same execution mode as Server with only a tighter latency setting.

Tested Azure Configurations

Platform	GPUs	Racks	Reported nodes
Azure ND GB200 v6	72	1	18
Azure ND GB200 v6	288	4	72
Azure ND GB300 v6	72	1	18

A full NVL72 rack contains 72 GPUs. Azure ND GB200 v6 and ND GB300 v6 are built on NVIDIA GB200 NVL72 and NVIDIA GB300 NVL72, respectively. Each node provides four GPUs and two Grace CPUs, with 18 nodes per rack. NVLink connects GPUs within the rack, while a separate scale-out network connects racks. In the four-rack GB200 submission, that network was Microsoft's Ethernet-based backend running Multipath Reliable Connection (MRC). These are distinct parts of the platform; the four-rack result tests capacity beyond a single NVLink domain. The single-rack GB200 and GB300 submissions used only NVLink.

The comparisons that follow hold model, scenario, GPU family and GPU count constant where applicable. They measure the entire submitted systems, including hardware, software and configuration. Matching hardware does not isolate the effect of software, networking or other potential optimization.

Interactive Performance Leadership

Azure ND GB300 v6 achieves **260,098 output tokens per second** in the Interactive scenario, the highest score among the five Available DeepSeek-R1 submissions in the analyzed results. Every system in this comparison uses 72 GPUs, with GPU generation identified in the chart.

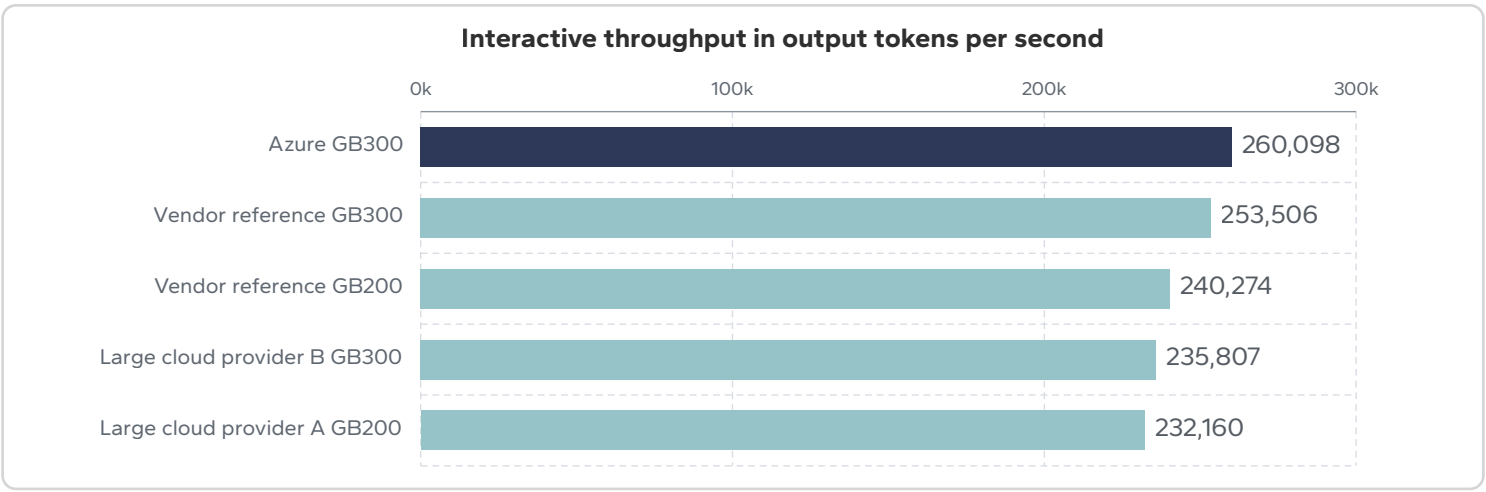


Figure 1: *All five Available DeepSeek-R1 Interactive submissions in the analyzed results, in output tokens per second. Each system has 72 GPUs. Provider descriptions are anonymized and GPU generation is shown.*

© 2026 Signal65. All rights reserved.

The closest cloud submission is a 72-GPU GB300 system at 235,807 tokens per second. Azure's **10.3% lead** therefore compares the same GPU generation and count. Against NVIDIA's 72-GPU GB300 reference, which records 253,506 tokens per second, Azure leads by **2.6%**.

The two GB200 entries provide additional category context. They achieve 240,274 and 232,160 tokens per second. Because those systems use an earlier GPU generation, their position in the chart should not be interpreted as a comparison that isolates cloud-platform performance.

Scaling to Multiple NVL72 Racks

Azure's four-rack GB200 submission is the strongest evidence in this result set of its ability to expand inference capacity beyond one rack. Moving from 72 to 288 GPUs increases Server throughput from **426,796 to 1,598,510 tokens per second** and Offline throughput from **542,075 to 2,091,190 tokens per second**.

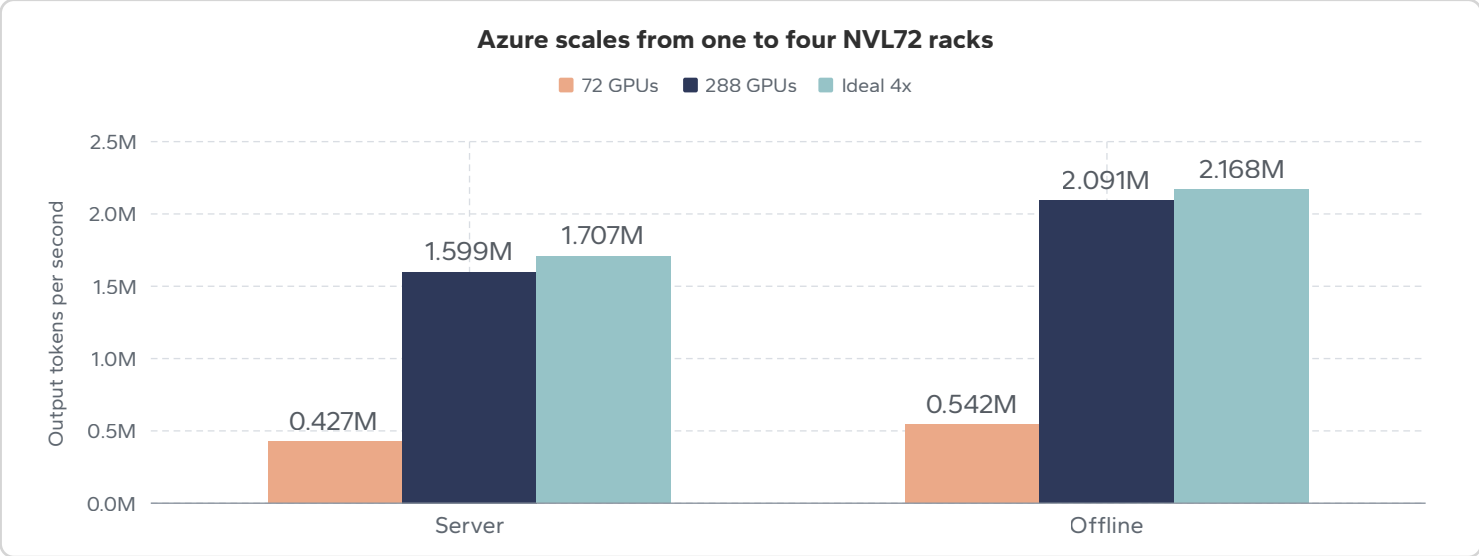


Figure 2: *Azure GB200 throughput at 72 and 288 GPUs, compared with four times the single-rack result. Both scenarios retain more than 93% of ideal scaling. Higher throughput is better.*

Measure	Server	Offline
Throughput increase	3.75x	3.86x
Scaling efficiency	93.6%	96.4%

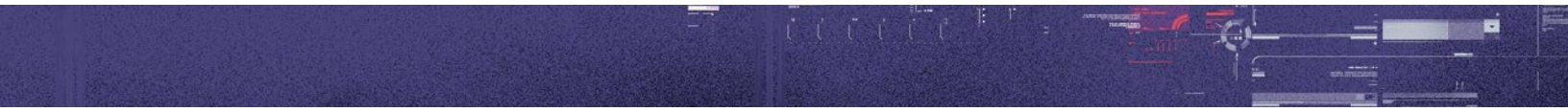
We calculated scaling efficiency as the four-rack throughput divided by four times the single-rack throughput. The figures show that a four-fold increase in GPUs delivers nearly four-fold aggregate output. The Server result is particularly relevant to online inference because that capacity is achieved while meeting the scenario's response-time requirements.

For an enterprise building a larger inference service, this provides a measured scale-out factor: Azure's GB200 results extend from one rack to a 72-node, 288-GPU four-rack system with a relatively small reduction in throughput per GPU. The evidence covers both a continuous-arrival workload and batch processing.

© 2026 Signal65. All rights reserved.

The results establish system-level scaling at the tested configurations. They do not isolate the scale-out network's contribution, nor establish that every request spans all four racks, nor demonstrate the same efficiency at larger scales.

In the tested four-rack GB200 configuration, multi-rack expansion used Microsoft's Ethernet-based scale-out backend running Multipath Reliable Connection (MRC), a RoCE-based transport co-developed by OpenAI, Microsoft and other industry partners for large AI clusters. Rather than binding each connection to a single network path, MRC spreads traffic across many paths at once, which is designed to make performance less sensitive to any single slow or failed link.



Azure Leads Available 72-GPU GB200 Submissions

At 72 GPUs, Azure ND GB200 v6 leads all four Available GB200 NVL72 submissions in both Server and Offline. It delivers **426,796 tokens per second in Server** and **542,075 tokens per second in Offline**. This comparison keeps GPU generation and counts consistent across the group.

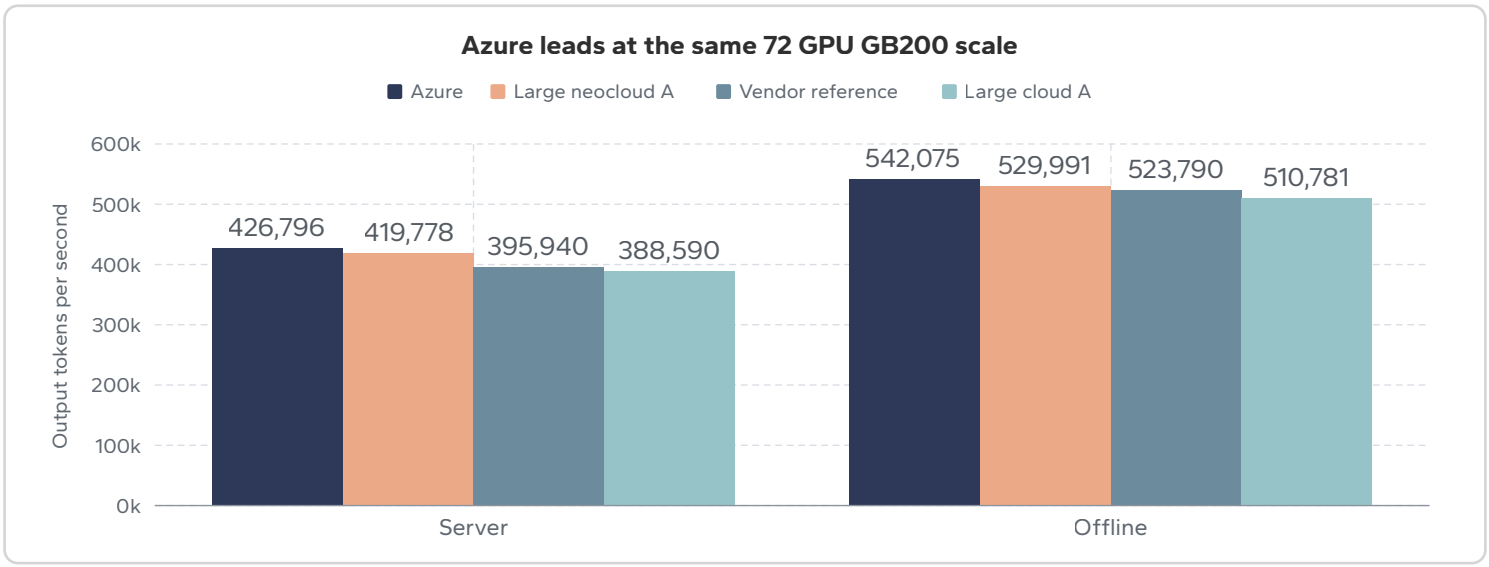


Figure 3: *Available 72-GPU GB200 NVL72 systems. The two scenarios are separate tests. Azure is highest in each, with the nearest cloud entry within 2.3%.*

The nearest cloud submission, Large neocloud provider A, records 419,778 tokens per second in Server and 529,991 in Offline. Azure's lead is **1.7% in Server and 2.3% in Offline**. These are modest margins, but they occur in both tested scenarios at the same accelerator count.

Against Large cloud provider A, Azure's advantage is **9.8% in Server and 6.1% in Offline**. Azure also exceeds the accelerator vendor reference by 7.8% and 3.5%, respectively. The appendix includes all four systems and their result IDs, so the comparison does not depend on selecting only the widest margin.

© 2026 Signal65. All rights reserved.

A Competitive Base for Expansion

The single-rack GB200 result provides the baseline for Azure's four-rack scaling calculation. Together, the results show that Azure starts from a leading position within this matched group and retains most of its per-GPU throughput when expanding to 288 GPUs.

This is evidence of effective performance from the complete Azure deployment. It does not isolate the contribution of any single component: the cloud and vendor-reference submissions are not a controlled experiment that changes only one variable. Software, tuning and other system details can also influence the differences.

GB300 Scale-Up Results

Azure's 72-GPU GB300 system delivers **589,391 tokens per second in Server** and **661,143 in Offline**. Relative to Azure's 72-GPU GB200 result in the same round, that is a **38.1% Server uplift** and a **22.0% Offline uplift**. Both comparisons use the same model and GPU count.

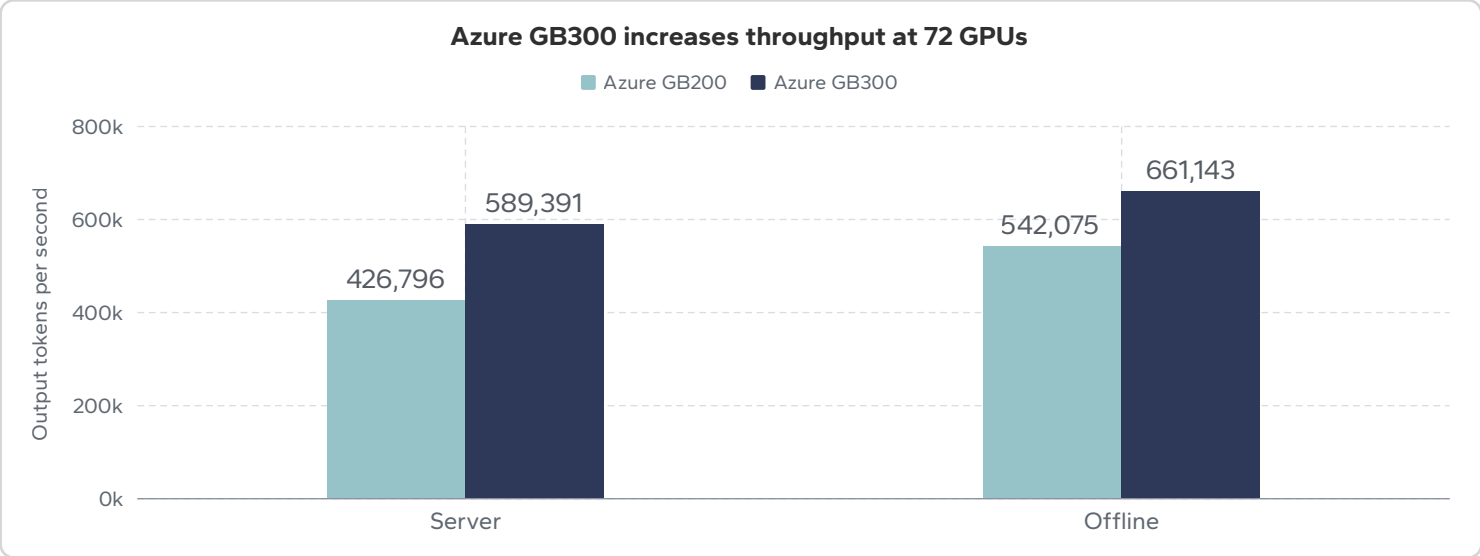


Figure 4: *Azure GB200 and GB300 at 72 GPUs in MLPerf Inference v6.1. Server throughput increases by 38.1% and Offline throughput by 22.0%.*

Microsoft's published GB300 specifications describe more HBM capacity per GPU than GB200. That difference provides relevant platform context. The benchmark comparison measures the complete submitted systems, and does not establish how much of the observed gain comes from memory, compute, networking or software.

Azure's Position Within the GB300 Group

Azure's strongest competitive result on GB300 is Interactive. Among the four Available 72-GPU GB300 systems, Azure's Server score is within 2.3% of the highest result of 603,023 tokens per second and its Offline score is within 4.2% of the highest result of 689,961. All four systems fall within a 5% band in both scenarios. The full peer group is retained in Appendix A.

© 2026 Signal65. All rights reserved.

These results show that Azure combines the highest Interactive score with Server and Offline throughput within a few percent of the corresponding GB300 leaders.

For deployment planning, the GB300 results establish higher single-rack throughput within Azure. The GB200 results separately establish four-rack scaling. A GB300 scale-out projection would require its own measurement; the two findings should not be combined into an untested multi-rack GB300 capacity claim.

Implications for Enterprise Deployment

Inference services often outgrow the capacity of a single NVL72 rack. Scaling across multiple racks requires balanced CPU, GPU, memory, networking and storage connectivity, along with the operational management to run it. Azure's results cover both ends of that range, with a single-rack platform for contained deployments and a measured four-rack system for larger ones. By offering a range of public cloud configurations, Azure can help enterprises scale workloads as demand changes.

Selecting Scenarios with Use Cases

For applications sensitive to response time, Azure's GB300 Interactive leadership is the most relevant starting point. It shows aggregate throughput under the benchmark's tight response time constraints. A production evaluation should preserve the application's prompt lengths, output lengths, arrival patterns and latency objectives when checking whether that advantage applies.

For online services with less restrictive response requirements, the Server results provide the closest benchmark reference. Azure demonstrates 589,391 tokens per second on a single GB300 rack and 1,598,510 on four GB200 racks. These are measured configurations that can anchor a capacity evaluation.

For asynchronous work, Offline is the relevant throughput measure. Azure's four-rack GB200 system exceeds two million output tokens per second while retaining 96.4% scaling efficiency. This is the strongest batch-capacity result in Azure's submitted set and a direct demonstration that its measured inference capacity extends beyond a single rack.

Applying Benchmark Results

Reported MLPerf results are useful for comparing competing offerings and relative performance, but they do not provide deployment-specific sizing guidance. Production sizing depends on factors including model settings, retrieval patterns, tool-call rates, request routing and other workload characteristics. These considerations are outside the DeepSeek-R1 throughput scores analyzed here. Evaluation teams should validate their requirements before translating benchmark output into user capacity, service-level commitments, or specific system configurations.

Azure demonstrates leading Interactive performance, leadership within the 72-GPU GB200 group, and higher single-rack throughput with GB300. The four-rack GB200 result adds an important measurement: **Azure sustains near-linear aggregate throughput growth from 72 to 288 GPUs in both Server and Offline.**

For infrastructure teams planning growth beyond one rack, the four-rack result is the most useful data point in Azure's submission set, because it replaces an assumption about how a strong single-rack score will scale with a direct measurement of near-linear scaling at four racks.

© 2026 Signal65. All rights reserved.

Appendix A: Selected Results

All entries below are MLPerf Inference v6.1 Closed Datacenter DeepSeek-R1 results in the Available category. Values are output tokens per second. A dash means no result is recorded for that scenario. These tables cover the comparison groups used in the paper, rather than every system in the overall dataset.

GB200 NVL72 at 72 GPUs

Provider	Result ID	Server	Offline	Interactive
Microsoft Azure	6.1-0009	426,796	542,075	—
Large neocloud provider A	6.1-0023	419,778	529,991	—
Vendor reference	6.1-0072	395,940	523,790	240,274
Large cloud provider A	6.1-0047	388,590	510,781	232,160

GB300 NVL72 at 72 GPUs

Provider	Result ID	Server	Offline	Interactive
Microsoft Azure	6.1-0010	589,391	661,143	260,098
Large neocloud provider B	6.1-0080	603,023	689,961	—
Vendor reference	6.1-0074	596,944	679,740	253,506
Large cloud provider B	6.1-0087	586,852	666,613	235,807

Azure GB200 NVL72 at 288 GPUs

Provider	Result ID	Server	Offline	Interactive
Microsoft Azure	6.1-0008	1,598,510	2,091,190	—

Figure 1 combines the three GB300 entries with Interactive scores and the two GB200 entries with Interactive scores. Neocloud entries without Interactive scores are not treated as zero. Preview systems are excluded from all leadership claims.

Comparison scope

Provider descriptions are consistent throughout the report and accompanying workbook. Result IDs preserve traceability to the source entries. Matching accelerator generation and count improves the comparison, but does not make the submissions identical in software, precision settings, host configuration or tuning.

© 2026 Signal65. All rights reserved.

Appendix B: Calculations and Sources

Headline calculations

Measure	Calculation	Result
Interactive lead over next cloud	260,098 / 235,807 − 1	10.3%
Interactive lead over GB300 reference	260,098 / 253,506 − 1	2.6%
GB200 lead over next cloud in Server	426,796 / 419,778 − 1	1.7%
GB200 lead over next cloud in Offline	542,075 / 529,991 − 1	2.3%
Azure GB300 Server uplift	589,391 / 426,796 − 1	38.1%
Azure GB300 Offline uplift	661,143 / 542,075 − 1	22.0%
GB200 Server scaling efficiency	1,598,510 / (4 × 426,796)	93.6%
GB200 Offline scaling efficiency	2,091,190 / (4 × 542,075)	96.4%

Scaling efficiency compares the measured 288-GPU result with four times Azure's measured 72-GPU result. It is also the fraction of single-rack throughput per GPU retained at four racks. Percentages are rounded to one decimal place.

Published Azure platform context

Specification	ND GB200 v6	ND GB300 v6
GPUs per node / full NVL72 rack	4 / 72	4 / 72

These are published product specifications from Microsoft, provided as architecture context.

© 2026 Signal65. All rights reserved.

Selected public sources

MLCommons MLPerf Inference v6.1 results Published results and submission artifacts. Result IDs used in this paper are listed in Appendix A.

MLCommons Inference rules DeepSeek-R1 scenario constraints and permitted optimizations.

Microsoft ND GB200 v6 specifications VM and interconnect architecture.

Microsoft ND GB300 v6 specifications VM and interconnect architecture.

Microsoft: Building resilient networks for AI supercomputers MRC scale-out network transport.

Sources reviewed September 16, 2026. MLPerf® Inference v6.1, Closed division, Datacenter, DeepSeek-R1, Interactive, Server and Offline scenarios, Available category. Results retrieved from mlcommons.org/benchmarks/inference-datacenter on September 16, 2026. Result IDs 6.1-0008, 6.1-0009, 6.1-0010, 6.1-0023, 6.1-0047, 6.1-0072, 6.1-0074, 6.1-0080 and 6.1-0087. Results verified by MLCommons Association. 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 strictly prohibited. See www.mlcommons.org for more information. The accompanying analysis workbook includes the broader result dataset, historical inputs and formula-linked comparisons.

© 2026 Signal65. All rights reserved.

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

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and lab services to many high-tech companies, including those mentioned in this paper.

ABOUT SIGNAL65

Signal65 exists to be a source of data in a world where technology markets and product landscapes create complex and distorted views of product truth. We strive to provide honest and comprehensive feedback and analysis for our clients in order for them to better understand their own competitive positioning and create optimal opportunities to market and message their devices and services.

IN PARTNERSHIP WITH

CONTACT INFORMATION
Signal65 | signal65.com

© 2026 Signal65. All rights reserved.