signal65

November 2026

MLPERF INFERENCE V6.1 | DEEPSEEK-R1

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

Azure ND GB300 v6, built on NVIDIA GB300 NVL72, posts the highest Interactive throughput among all Available submissions, and Azure ND GB200 v6 scales near-linearly from one rack to four.

10.3%

Interactive throughput lead over the next cloud submission, at 260,098 output tokens per second on 72 GPUs

Highest Interactive throughput of all five Available submissions

Every system uses 72 GPUs. Azure leads NVIDIA's own GB300 reference by 2.6%. Output tokens per second, higher is better.



93.6%

Server scaling efficiency from 72 to 288 GPUs (96.4% in Offline)

2.09M

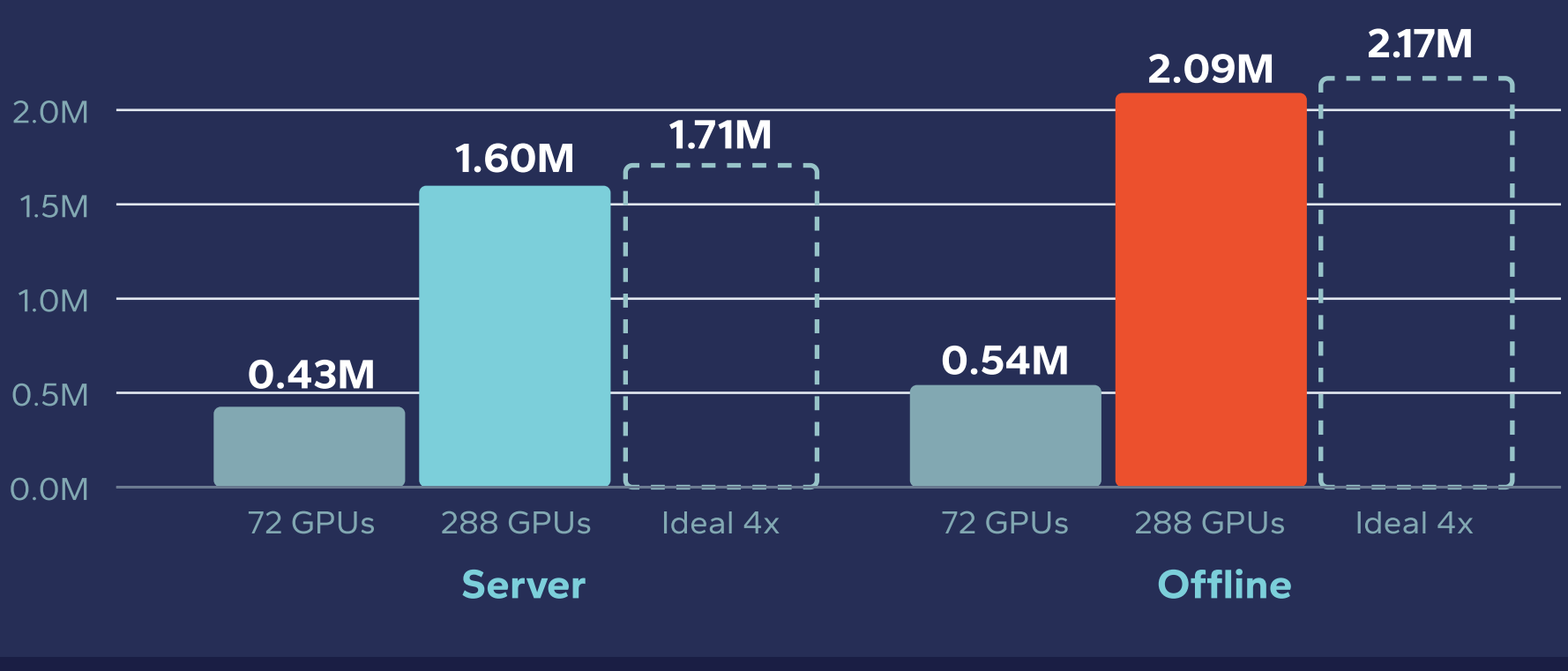
Offline output tokens per second on four ND GB200 v6 racks

38.1%

Server throughput gain from GB200 to GB300 at 72 GPUs (22.0% in Offline)

Near-linear scaling beyond a single NVL72 rack

Four racks deliver 3.75x the Server throughput and 3.86x the Offline throughput of one rack. Output tokens per second.



The four-rack run used Microsoft's Ethernet-based Multipath Reliable Connection (MRC) scale-out network.

Only cloud

In its first DeepSeek-R1 MLPerf submission, Azure is the only cloud provider with a multi-rack NVL72 result, so multi-rack capacity is measured directly.

Leads all 4

Azure ND GB200 v6 leads all four Available 72-GPU GB200 systems in both Server and Offline, by 1.7% and 2.3% over the closest cloud submission.

Which result fits which workload

The three MLPerf scenarios are separate tests with different latency limits.

Interactive
TTFT 1.5 s | TPOT 15 ms
Conversational tools and coding assistants where users wait on output

Server
TTFT 2.0 s | TPOT 80 ms
Online inference services with less restrictive response targets

Offline
No latency SLO
Batch and asynchronous work where total throughput matters most

Commissioned by:

Microsoft

Source: Signal65 Lab Insight, "Microsoft Azure Leads DeepSeek-R1 Interactive Inference in MLPerf v6.1," November 2026. Full report at signal65.com

Analysis of MLPerf Inference v6.1 Closed Datacenter DeepSeek-R1 results. Comparisons hold model, scenario, GPU generation and GPU count constant. Competing submitters are anonymized. Scaling efficiency and generational gains are derived and are not verified by MLCommons.

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.