

AI Performance: Industry-leading servers and benchmarks

Joshua Basseby – HPE HPC & AI Performance Team

5th May 2026.

What is AI performance to HPE?

AI performance at HPE means the demonstration of AI workloads under extreme circumstances to prove the quality of HPE systems and solutions. HPE strives to show the world, our customers, and ourselves what is possible when you focus on excellence, then show others how to do the same.

It isn't enough to understand that a technology 'just works', instead ask 'how well does it work?'



What does performance tell you? - Top500

- HPC benchmarks designed for massive scale-out and high-bandwidth performance
 - ✓ **Thousands** of nodes, **Millions** of cores, **MegaWatts** of power, **ExaFLOPs** of performance
 - ✓ **LINPACK**: Reflects system performance by solving a standardized dense system of linear equations
 - ✓ Double-precision FP64

- HPE systems dominate Top500 (#1-3, 6 in top 10, 13 in top 25 clusters)¹

- ✓ **HPE Cray EX (x86_64 CPUs, AMD GPUs, HPE Slingshot)** #1-3 top clusters

Nodes	Cores	Power	Performance	Rank
10,000+	11,000,000+	29 MegaWatts	1.8 ExaFLOPs	#1

- ✓ **HPE Cray XD670 (x84_64 CPUs, H200, IB)** #16 (2nd APJ cluster), #78 (top Canadian cluster)

Nodes	Cores	Power	Performance	Rank
400+	479,000+	3.6 MegaWatts	145 PetaFLOPs	#16

¹ Based on Top500 - [November 2025](#)

What is MLPerf and Why do we use it?

- MLPerf is a set of open industry-standard benchmarks, results, and reference architectures
- MLPerf measures quality and performance to **prove and improve: AI accelerators, server hardware, software, and AI models** through...
 - ✓ **Benchmarking** for HPC, Storage, AI Training, AI Inference, AI Safety, & more
 - ✓ **Datasets** that are open, large-scale, and diverse
 - ✓ **Research** open collaboration and support with the research community helps accelerate and democratize scientific discovery, derive new insights for new breakthroughs in AI with **reference code and systems**
- Industry-wide consortium to **compare like-to-like performance** on standard benchmarks (**FP16-FP4**)

BY THE NUMBERS

Accelerating AI Innovation

At MLCommons, we democratize AI through open, state-of-the-art industry-standard benchmarks and data tooling to measure quality, performance, and risk.

125 +

MLCommons Members and Affiliates

10

Benchmark Suites

89.7 k +

MLPerf Performance Results to-date

700 k

Datasets using the Croissant metadata vocabulary

2025 HPE XD-series performance

- Focus on Cray XD670 with NVIDIA Hopper (H100-SXM and H200-SXM)
 - ✓ **Top performer:** MLPerf Inference and Training 2025
 - ✓ **Innovation:** HPE only company to publish Llama3.1-8B pretraining results on H200
 - ✓ MLPerf Training [v5.0](#) and [v5.1](#): #1 in 13 of 19 scenarios (68%)¹
 - ✓ MLPerf Inference [v5.0](#) and [v5.1](#): #1 in 36 of 83 scenarios (43%)¹
 - ✓ Reference architectures and reproducible code/results
- External partners and customer use HPE XD-series for performance publications...
 - ✓ Krai, Ltd. (MLPerf partner, **H200 fastest-ever Llama2 fine-tuning** in MLPerf Training v5.1)¹
 - ✓ [Sovereign AI Factory](#) (#78 of Top500, **fastest cluster in Canada**)



Training

#1: 68%

Inference

#1: 43%



HPE ProLiant DL380a Gen12:
Leadership with up to 10x PCIe-
based GPUs on AI inference
workloads

Inference

#1: 71%

- ✓ MLPerf Inference [v5.0](#) and [v5.1](#): #1 in 32 of 45 scenarios (71%)¹
[Read more: https://community.hpe.com/t5/ai-unlocked/hpe-delivers-several-world-records-in-latest-mlperf-inference/ba-p/7255374](https://community.hpe.com/t5/ai-unlocked/hpe-delivers-several-world-records-in-latest-mlperf-inference/ba-p/7255374)

¹ When compared to similar systems and GPUs in MLCommons MLPerf [Training](#) and [Inference](#)

XD685 DLC performance B200 vs H200

- Focus for 2026 performance studies
- 8x NVIDIA B200
 - ✓ MLPerf Inference v6.0: Llama3.1-8B vs 8x H200¹



- ✓ MLPerf Inference v6.0: Llama2-70B 99.9 vs 8x H200¹



- ✓ MLPerf Inference v6.0: Mixtral-8x7B vs 8x H200¹



 **High throughput**

 **Low Latency**

 **Cooling Efficiency**

¹ MLPerf Inference v6.0 "HPE XD685 with 8x NVIDIA B200" (submission ID 6.0-0049) compared to MLPerf Inference v5.1 "HPE XD670 with 8x NVIDIA H200" (submission ID: 5.1-0049), <https://mlcommons.org/benchmarks/inference-datacenter/>



Thank You

