Intel Arc Pro B-Series GPUs and Xeon 6 Shine in MLPerf Inference v5.1

SEPTEMBER 9, 2025 Published

Artificial Intelligence

MLPerf v5.1 benchmarks showcase Intel Xeon 6 and Arc Pro B-Series GPUs delivering powerful, low-latency AI inference for workstations and edge systems.

Share

What's New: Today, MLCommons released its latest MLPerf Inference v5.1 benchmarks, showcasing results across 6 key benchmarks for Intel's GPU Systems featuring Intel® Xeon® with P-cores and Intel® Arc™ Pro B60 graphics, inference workstations code-named Project Battlematrix. In Llama 8B, Intel Arc Pro B60 performance per dollar advantages of up to 1.25x and up to 4x compared to NVIDIA RTX Pro 6000 and L40S respectively.[1] The results underscore the performance and accessibility of an all-Intel platform that addresses emerging AI inference workloads across high-end workstations and edge applications.

> "The MLPerf v5.1 results are a powerful validation of Intel's GPU and AI strategy. Our Arc Pro B-Series GPUs with a new inference-optimized software stack let developers and enterprises develop and deploy AI powered applications with inference workstations that are powerful, simple to set up, accessibly priced, and scalable."
>
> - Lisa Pearce, Intel corporate vice president and general manager of Software, GPU and NPU IP Group

Why It Matters: Until now, limited options existed for professionals who prioritized platforms capable of delivering high inference performance without compromising data privacy or incurring heavy subscription costs tied to proprietary AI models, but required capabilities to deploy large language models (LLMs).

These new Intel GPU Systems, code-named Project Battlematrix, are designed to meet the needs of modern AI inference and provide an all-in-one inference platform combining full-stack validated hardware and software.

Intel GPU systems aim to simplify the adoption and ease of use with a new containerized solution built for Linux environments, optimized to deliver incredible inference performance with multi-GPU scaling and PCIe P2P data transfers, and designed to include enterprise-class reliability and manageability features such as ECC, SRIOV, telemetry and remote firmware updates.

CPUs continue to play a vital role in AI systems. As the orchestration hub, the CPU handles preprocessing, transmission, and overall system coordination. Intel sustained improvements in CPU-based AI performance over the past four years have established Intel Xeon as the preferred CPU for hosting and managing AI workloads in GPU-powered systems.

Intel also remains the only vendor submitting server CPU results to MLPerf, demonstrating leadership and a deep commitment to accelerating AI inference capabilities across both compute and accelerator architectures. Notably, Intel Xeon 6 with P-cores achieved 1.9x performance improvement gen-over-gen in MLPerf Inference v5.1

More Context: MLPerf Inference v5.1 Results | Project Battlematrix

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. Visit MLCommons for more details. No product or component can be absolutely secure.

[1]: Cost estimates are based on internal estimates from public and partner sourcing for a configuration using Intel Xeon w7-2475x, 4x Arc Pro B60 Dual GPU cards and 2 memory sticks of 64GB DDR5 5600MHz memory as of September 2025.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Artificial Intelligence | Intel Xeon | Intel® Arc™ Pro B60 | MLPerf

Copy text from article