GPU SMITH
COMPUTE ENGINEERING

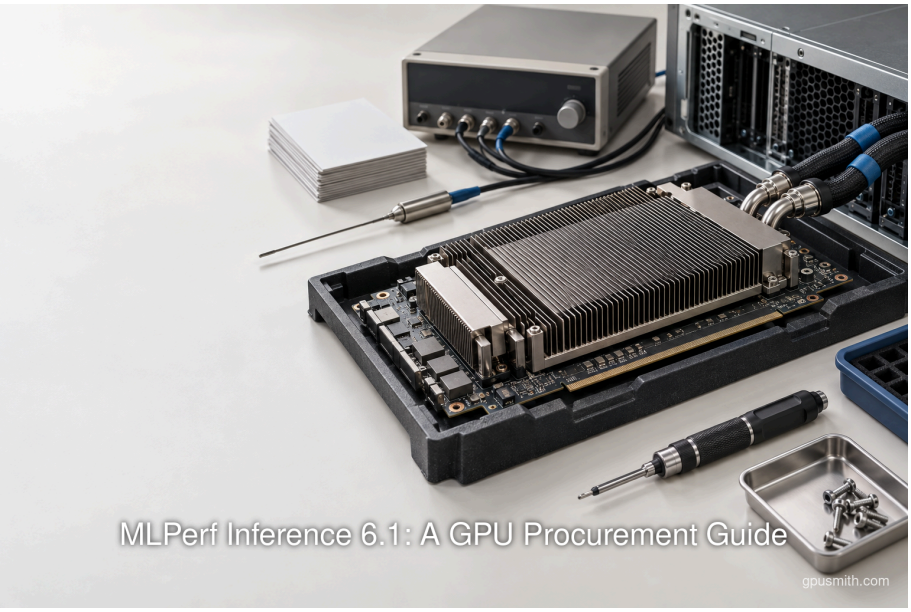
GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

MLPerf Inference 6.1: A GPU Procurement Guide

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-09-19 · mlperf inference 6.1 · gpu procurement · mlperf benchmarks · inference infrastructure · gpu performance · server benchmarking · ai infrastructure · performance per watt · private ai

Original article: https://gpusmith.com/articles/en/mlperf-inference-6-1-gpu-procurement



In this article

Executive Summary

MLPerf Inference 6.1, published on **September 16, 2026**, is useful for GPU procurement when treated as a structured evidence set, not a universal leaderboard. The release covers **10 datacenter and 6 edge benchmarks** (mlcommons.org) and **120 systems** (mlcommons.org). It adds end-to-end retrieval-augmented generation (RAG), an edge agentic test, and permitted speculative decoding for GPT-OSS-120B Interactive (mlcommons.org). More than half of submitters used the API-centric harness (mlcommons.org). These changes improve relevance, but they also create more ways to compare unlike rows.

The procurement rule is simple: compare only a cohort with the same release, suite, division, benchmark, accuracy target, scenario, availability class, unit, and applicable quality or latency constraint. **Closed** is the foundation for like-for-like comparison because its model must remain mathematically equivalent (mlcommons.org). **Open** can disclose useful techniques, but it permits arbitrary preprocessing, postprocessing, and model choices (github.com). Likewise, Offline sends all samples in one query (github.com), while Server and Interactive model Poisson request arrivals (github.com). Their outputs answer different purchasing questions.

The worked scale example illustrates the discipline. A Closed, available Qwen3-VL Offline row reports **58.395875 samples/s on four GB200 accelerators** (raw.githubusercontent.com); the 16-accelerator family member reports **230.088793 samples/s** (raw.githubusercontent.com). That yields **98.50% descriptive scale efficiency**, but it does not prove that any buyer will reproduce the result or that one GPU will deliver the quotient.

MLPerf does not provide price, lead time, production utilization, energy tariff, or facility cost. Buyers should attach dated quotes, metered power, support scope, and their own traffic trace. The v6.1 summary contains no power result for the sampled rows, so GPU thermal design power must not be substituted for measured system energy. A shortlist should therefore end in a buyer-owned pilot: clone exact artifacts, freeze hardware and software, run accuracy and performance, then replay the buyer's prompts and service-level objective (SLO). GPU Smith's documented approach similarly derives topology and node count from workload and throughput requirements (gpusmith.com) and validates against written acceptance criteria (gpusmith.com).

Introduction and Background

An MLPerf result is a measurement of a declared system under a defined workload. It is not a GPU specification, a price sheet, or a forecast of an application's service level. MLCommons describes the suite as designed to measure machine-learning model performance (docs.mlcommons.org), while the public v6.1 repository contains both results and submitted code (github.com). The distinction matters for enterprise AI infrastructure buyers, platform owners, performance engineers, finance teams, and investors evaluating private inference.

The correct first question is not, "Which accelerator won?" It is, **"Which benchmark and scenario resemble the service being purchased?"** A batch pipeline with a full queue may map to Offline. A request-serving endpoint may map to Server. An interactive large-language-model (LLM) service requires attention to time to first token (TTFT), time per output token (TPOT), and throughput together. MLPerf separately collects TTFT for the first generated token (raw.githubusercontent.com). A multi-stage knowledge system may map more closely to the new RAG workload, which reports documents per second for ingestion and tasks per second for question answering (mlcommons.org).

Where none of these models, request distributions, context lengths, quality targets, or latency limits resembles the intended service, the benchmark cannot justify a hardware conclusion. It can still identify systems for a lab pilot. This report provides the worksheet that connects the public evidence to that pilot without treating cross-row ratios as procurement facts.

Key Changes in MLPerf Inference 6.1

New end-to-end and agentic workloads

The **end-to-end RAG** test measures a pipeline of multiple models for question answering (mlcommons.org). It includes embedding, retrieval, reranking, and LLM reasoning. It exposes two independent workloads: corpus ingestion and question answering against a prebuilt vector database (mlcommons.org). The first release focuses on Offline rather than Server (mlcommons.org). Consequently, it informs throughput sizing for saturated pipelines but does not directly establish an online RAG latency SLO.

The **Edge Agentic** benchmark moves from single-shot inference toward multi-turn work (mlcommons.org). It combines an edge model, quantization, and single-stream coding workload (mlcommons.org). Buyers should not transfer an edge single-stream result to a datacenter concurrency forecast.

Optimization and harness changes

GPT-OSS-120B Interactive now permits speculative decoding, but the rule is not an unrestricted license. Speculative decoding is allowed only for specific benchmark and scenario combinations (raw.githubusercontent.com), and implementations must use the reference multi-token prediction head (raw.githubusercontent.com). A vendor percentage that mixes a speculative Interactive entry with a non-speculative or different-scenario baseline is not a valid procurement ratio.

The API-centric endpoint harness uses a client/server architecture over standard APIs (mlcommons.org). Its artifacts differ from traditional LoadGen submissions: endpoint results use structured JSON and YAML files (github.com), including `result_summary.json` and `accuracy_results.json` (github.com). Procurement reviewers should therefore expect different filenames without treating that difference as missing evidence.

New hardware and availability labels

MLCommons classifies **AMD Instinct MI350P** and **Intel Arc Pro B70** as available, while **NVIDIA Rubin** and **Vera Rubin NVL72** are preview in this round (mlcommons.org) (mlcommons.org). In MLPerf terminology, Available means all components can be purchased or rented (docs.mlcommons.org); Preview means expected availability in the next submission round (docs.mlcommons.org). It is a benchmark classification, not a quotation, allocation, support term, or promised delivery date.

Building a Valid Comparison Cohort

Every candidate row should pass the same filter before arithmetic begins. *Table 1* is the minimum comparable-cohort worksheet.

Field	Include only when	Exclude or separate when	Procurement reason
Release and suite	v6.1 and the same Datacenter or Edge suite	Different versions or suites	Workload definitions and rules can change.
Division	Closed with Closed, or Open with Open	Closed mixed with Open	Closed preserves mathematical equivalence; Open permits broader changes.
Benchmark and target	Same model, workload, accuracy target, and quality constraint	Different models, target suffixes, or achieved accuracy	Throughput is conditional on acceptable output quality.
Scenario	Same Offline, Server, Interactive, SingleStream, or MultiStream row	Cross-scenario ratios	Arrival process and latency obligations differ.
Availability	Same Available cohort for near-term procurement	Preview or RDI ranked as deliverable	Benchmark status is not delivery evidence.
Unit	Same tokens/s, samples/s, queries/s, tasks/s, or documents/s	Unlike output units	A numeric ratio across units has no operational meaning.
Constraint	Same TTFT, TPOT, percentile latency, duration, and valid flag	Different latency or quality limits	Throughput is only valid inside the row's constraints.
System scale	Full accelerator, node, CPU, memory, network, and software context retained	Accelerator-only labels	A result belongs to the complete system under test.

The table prevents a common error: sorting visually adjacent results that answer different questions. MLCommons supports Offline, Server, Interactive, SingleStream, and MultiStream scenarios (mlcommons.org). A comparison is defensible only after every row passes every relevant field.

Closed versus Open

Closed permits calibration for quantization but no retraining (github.com). This makes it the preferred division for a procurement ranking, while still requiring inspection of precision, kernels, runtime, and scenario settings. **Open** reports achieved accuracy rather than Closed accuracy constraints (github.com). It can reveal techniques worth testing, but it must remain a separate analytical cohort.

Lambda's v6.1 explanation is a useful illustration: its Open agentic entry kept the reference workload but swapped the served model (lambda.ai). Lambda expressly says the result is not an apples-to-apples comparison (lambda.ai). That is an optimization lead, not a price-performance ranking input.

Server, Offline, and Interactive

- **Offline:** asks how quickly a fully available queue can be completed. It favors sustained batch throughput.
- **Server:** tests a target query rate under random arrivals and applicable latency constraints. It is closer to concurrent request serving.
- **Interactive:** adds user-visible generation constraints. It should be assessed through TTFT, TPOT, and throughput, not throughput alone.
- **Do not substitute:** Offline throughput for Server capacity, or aggregate throughput for user-visible responsiveness.
- **Map the trace:** compare the benchmark's input and output lengths, arrival pattern, concurrency, and quality threshold with the buyer's traffic.

Server throughput should normally be expected to differ from Offline throughput because the request process and latency obligations change. This is not inefficiency in the colloquial sense. It is the cost of satisfying a different service condition.

Reading the System, Not Just the Accelerator

The repository's system-description JSON combines hardware and software information (raw.githubusercontent.com). Each shortlisted row should be expanded into an evidence record:

- **Identity:** submitter, system name, repository result path, division, suite, availability, and validity.
- **Workload:** benchmark, scenario, accuracy target, result, unit, and latency or quality constraints.
- **Accelerators:** exact model, count, memory, interconnect, and configured power settings.
- **Host:** CPU model and count, DRAM capacity, NUMA layout, storage, and operating system.
- **Fabric:** network interface, link rate, switch topology, and whether traffic crosses nodes or regions.
- **Software:** driver, runtime, framework, inference server, quantization, kernel, model commit, and harness.

- **Thermal design:** air or liquid cooling, rack density, facility water and power assumptions, and measured run conditions.
- **Artifacts:** summary row, system JSON, measurements, config, code, accuracy output, performance output, audit files, and supplemental statement.

Hardware specifications show why the system boundary matters. AMD lists **MI350P** with **600 W maximum board power**, configurable to 450 W (amd.com), while its v6.1 technical material lists **144 GB HBM3E and 4 TB/s** bandwidth (amd.com). The same source describes the part as a **PCIe 5.0 dual-slot** card (amd.com). These are component specifications, not measured system energy or application throughput.

At rack scale, NVIDIA's **GB300 NVL72** reference architecture is liquid cooled (docs.nvidia.com), specifies **130 TB/s aggregate NVSwitch bandwidth** (docs.nvidia.com), and can require up to **142 kW** (docs.nvidia.com). Its ConnectX-8 design offers up to **800 Gb/s per SuperNIC** (docs.nvidia.com). A per-accelerator quotient that omits this topology and facility context is descriptive bookkeeping, not a deployable design.

OEM configuration is equally material. Dell documents its **10U PowerEdge XE9785** with eight B300 SXM6 GPUs at **270 GB and 1,100 W each** (dell.com), while Supermicro lists a direct-liquid-cooled four-rack-unit system with eight MI355X accelerators (supermicro.com). HPE separately lists an **MI350P PCIe accelerator** in its portfolio (hpe.com). Intel announced Arc Pro B70 availability beginning **March 25, 2026** (intel.com). These facts change rack count, cooling plant, and bill of materials even when the GPU label appears similar.

Software creates another boundary. Dell disclosed TensorRT 10.14, CUDA 13.1, cuDNN 9.17, and TensorRT-LLM for B300 submissions (infohub.delltechnologies.com). Dell also describes an MI350P recipe using MXFP4 W4A4 weights and an FP8 key-value cache (infohub.delltechnologies.com). Intel identifies PyTorch vLLM on Ubuntu 24.04 for its Arc Pro B70 disclosure (intel.com), and reports a four-GPU B70 node with **128 GB of VRAM** (intel.com). Neither stack should be assumed portable to another vendor or release.

Implementation Considerations and Process Changes

Trace a headline to artifacts

A procurement analyst should be able to walk from a slide to a public result without relying on the slide's arithmetic:

1. **Record the claim exactly.** Capture the stated ratio, benchmark, scenario, system scale, and whether MLCommons verified it.
2. **Locate the summary row.** Confirm submitter, system name, division, availability, benchmark, scenario, accuracy, result, and unit.
3. **Open the system JSON.** Verify accelerator count, CPU, memory, node count, network, storage, cooling, and software.
4. **Open the configuration.** Confirm precision, quantization, batch controls, speculative decoding, TTFT, TPOT, and target query rate.

5. **Open measurements and logs.** Confirm the performance run, accuracy run, valid flag, audit output, and any power workflow.
6. **Inspect code and README.** Determine whether the artifacts are sufficient to build and reproduce the implementation.
7. **Read supplemental material cautiously.** Treat submitter statements as context, not replacements for audited fields.
8. **Freeze the cohort.** Only after all fields match should ratios or normalized outputs be calculated.

MLPerf requires one LoadGen validation run for each submitted performance result (github.com), and accuracy and performance modes must execute the same code (github.com). Submission source must be sufficient to reproduce results (github.com). These controls strengthen evidence, but a buyer still needs to reproduce it on the intended platform.

Vendor materials can help locate differences. AMD provides step-by-step reproduction instructions for its measurements (rocm.blogs.amd.com). NVIDIA labels some later figures as post-submission and not yet verified by MLCommons (blogs.nvidia.com). Such labeling should flow into the evidence record and prevent an unofficial figure from entering an audited ranking.

Convert the shortlist into a pilot

- **Clone by commit:** save the final v6.1 repository commit and exact result paths.
- **Freeze inputs:** record model weights, tokenizer, dataset, precision, calibration, and accuracy target.
- **Freeze hardware:** record part numbers, firmware, CPU, memory, storage, NICs, switches, cooling, and topology.
- **Freeze software:** record OS, kernel, drivers, runtime, framework, server, compiler, containers, and orchestration.
- **Run accuracy first:** reject a configuration that misses the required benchmark or application quality threshold.
- **Run benchmark performance:** reproduce the applicable scenario and retain full logs.
- **Replay the buyer trace:** use real prompt lengths, output lengths, arrival distribution, concurrency, model mix, and retrieval corpus.
- **Measure service behavior:** capture throughput, TTFT, TPOT, tail latency, errors, recovery, and sustained utilization.
- **Measure facilities:** meter node or rack input power and validate cooling behavior at steady load.
- **Apply acceptance thresholds:** use buyer-supplied limits for quality, latency, availability, power, acoustics, and operations.
- **Reconcile economics:** attach dated hardware, network, support, installation, colocation, power, and staffing quotes.

- **Archive evidence:** preserve manifests, commands, logs, telemetry, deviations, and signed acceptance results.

NIST warns that common software performance practices can yield results that are plausible but wrong (nist.gov). SPEC's current run rules likewise state that energy metrics must not be estimated (spec.org). The pilot is therefore an evidence-transfer exercise, not a ceremonial rerun.

Data Analysis and Evidence

Table 2 shows how a few exact repository rows should be recorded. It is an evidence sample, not a cross-vendor leaderboard.

Evidence row	Comparable fields	Published system result	Descriptive calculation	Artifact interpretation
AMD 8x MI355X	Closed, Available, GPT-OSS-120B, Offline, base row	**121,818 tokens/s** (raw.githubusercontent.com)	**15,227.25 tokens/s per accelerator**	The configuration identifies the model path; the quotient is not a single-GPU prediction (raw.githubusercontent.com).
Crusoe 512x MI355X	Closed, Available, GPT-OSS-120B, Offline, base row	**5,749,440 tokens/s** (raw.githubusercontent.com)	**11,229.38 tokens/s per accelerator**	64 nodes and 512 single-GPU replicas make it materially different from the eight-GPU row (raw.githubusercontent.com).
Crusoe 4x GB200	Closed, Available, Qwen3-VL, Offline	**58.395875 samples/s** (raw.githubusercontent.com)	**14.598969 samples/s per accelerator**	Clean smaller member of the same submitted family.
Crusoe 16x GB200	Closed, Available, Qwen3-VL, Offline	**230.088793 samples/s** (raw.githubusercontent.com)	**14.380550 samples/s per accelerator**	Four-node member using the same declared Dynamo/vLLM framework string (raw.githubusercontent.com).
Dell MangoBoost 8x MI355X	Closed, Available, GPT-OSS-120B, Interactive	**29,585.5 tokens/s** (raw.githubusercontent.com)	**3,698.19 tokens/s per accelerator**	Interactive has tighter TTFT and TPOT constraints than its Server run (raw.githubusercontent.com).

The 4x to 16x Qwen3-VL pair has a scale factor of four. Its descriptive efficiency is `230.088792881784 / (58.395875238853066 × 4) = 0.9850`, or **98.50%**. This supports a narrow statement: the published larger configuration delivered 98.50% of ideal linear scaling relative to that smaller published configuration. It does not isolate networking as the cause of the 1.50% gap, prove production scaling, or authorize comparison with the Interactive rows.

The eight-GPU and 512-GPU MI355X quotients show why per-accelerator output should not become a ranking by itself. Their topology and replica strategies differ, and their results should not be compressed into a universal summary score.

Table 3 turns benchmark output into a buyer-owned financial and facility model. Blank fields are deliberate.

Buyer metric	Formula	Required evidence	Do not use
Validated output per accelerator	Published system result / published accelerator count	Exact row and system JSON	GPU name inferred from a slide
Same-family scale efficiency	Large result / (small result × accelerator-count ratio)	Matched scenario, quality, software, topology notes	Cross-vendor causal claim
Price-performance	Locally validated SLO-relevant result / dated quoted system cost	Quote with included hardware, network, support, tax, installation	Estimated street price
Energy per useful output	Metered joules for exact valid run / accepted outputs	Same-run power and performance logs	GPU thermal design power
Annual energy	Metered average kW × buyer duty hours	Measured utilization range and tariff	Nameplate power × 8,760 hours without qualification
Rack requirement	Quoted nodes / validated nodes per rack	OEM weight, power, cooling, and usable facility capacity	Nominal U-space alone
Facility cost range	Electrical, cooling, network, construction, and operations quotes	Site survey and uncertainty bounds	Benchmark result as a TCO proxy

MLPerf power policy includes all components sensitized by LoadGen (github.com) and requires power and performance from the same run (github.com). The selected v6.1 rows have no published power result, so this article does not calculate energy per token. That absence is a pilot requirement, not an invitation to substitute TDP.

Total cost should also extend beyond purchase price. U.S. Department of Energy guidance tells buyers to look beyond initial cost to total cost of ownership (energy.gov). Berkeley Lab's projected **325 to 580 TWh** range for U.S. data centers in 2028 illustrates the uncertainty created by equipment, utilization, and cooling assumptions (eta-publications.lbl.gov). The report itself notes that limited data availability constrains analysis (eta-publications.lbl.gov). NVIDIA also cautions that an illustrative average rack heat load cannot replace precise capacity planning (docs.nvidia.com). Facility inputs should therefore retain ranges.

Implications and Future Directions

The transition toward endpoints, RAG, and agentic work makes system-level benchmarking more useful, but also increases dependence on pipeline architecture. Intel says its RAG submission divided work between CPU and GPUs (intel.com). NVIDIA says its preview Vera Rubin entries used disaggregated serving with separate prefill and decode plus expert parallelism (blogs.nvidia.com). CoreWeave explicitly labels a reported per-GPU figure as not verified by MLCommons (coreweave.com), showing why verification status belongs in the evidence record. Future procurement will increasingly evaluate orchestration and networking alongside silicon.

Three practices follow:

- **Preserve raw evidence.** Save result paths, system descriptions, configs, logs, code, and commit identifiers in the RFP file.

- **Separate eligibility from scoring.** Availability, support, security, facility fit, and reproducibility are gates. Performance and economics are scored only after those gates pass.
- **Demand dated buyer evidence.** Quotes, lead times, warranty, support, metered power, and installation constraints belong beside the benchmark row.
- **Use preview for planning only.** A preview result can justify technical evaluation, not a delivery assertion or booked depreciation schedule.
- **Keep Open as a laboratory.** Techniques from Open may improve a later pilot, but should not enter a Closed price-performance ranking unchanged.

MLPerf's process provides meaningful safeguards: each submitter must review at least one other submission (github.com), and a review committee oversees publication (github.com). Those controls make v6.1 a strong shortlist input. They do not replace commercial diligence or application validation.

Frequently Asked Questions (FAQs)

How should buyers interpret MLPerf Inference 6.1 results?

Start with the workload, then filter to identical suite, division, benchmark, accuracy target, scenario, availability class, unit, and constraints. Read the system and configuration files before calculating ratios. Treat the result as evidence for selecting a pilot, not proof of an application SLO.

How can GPUs be compared using MLPerf?

Compare complete systems in a valid cohort. GPU labels alone omit host CPUs, memory, interconnect, node count, cooling, runtime, quantization, and inference software. When accelerator counts are verified, divide system output by count only as a labeled descriptive normalization. Do not predict a single-device result from it.

What is the difference between Server and Offline results?

Offline sends all samples at once. Server evaluates query rates under random arrivals and latency obligations. Aggregate Offline throughput answers a batch-capacity question; Server throughput answers a constrained request-serving question. Neither substitutes for Interactive TTFT and TPOT.

Can MLPerf show performance per watt?

Only when a valid power result exists for the exact run. Power policy requires full-system measurement and same-run performance. For the sampled v6.1 rows, no power result is published. A GPU's board-power specification is not measured system energy.

Does MLPerf provide GPU price-performance or TCO?

No. Add a dated quote with inclusions and exclusions to a locally validated result relevant to the SLO. Model support, networking, storage, installation, colocation, power, cooling, staffing, utilization, financing, and risk separately. Never insert an estimated street price into an audited comparison.

Which v6.1 submissions deserve an RFP shortlist?

Rows deserve a shortlist when their workload matches, cohort is valid, system is commercially eligible for the procurement date, artifacts are complete, facility requirements fit, and the supplier can support a reproduction pilot. A Preview entry may enter a technology watchlist, but not an Available-only delivery ranking.

What are MLPerf's limitations for GPU buyers?

MLPerf measures declared systems on standardized workloads. It does not supply the buyer's price, lead time, support scope, facility cost, production utilization, application quality, or traffic distribution. Its public result can establish a reproducible technical baseline, but it cannot establish total cost of ownership or prove a private inference SLO without local evidence.

What are the main red flags in vendor proposals?

- **Cherry-picked scenario:** a percentage combines Offline, Server, or Interactive without disclosure.
- **Mixed divisions:** Open is ranked against Closed as if the model and quality conditions were identical.
- **Preview sold as delivery:** benchmark availability status is treated as a promised ship date.
- **Aggregate record without scale:** accelerator count, nodes, fabric, or cooling is omitted.
- **Per-GPU arithmetic across unlike systems:** topology and replica strategy are ignored.
- **No result path:** a percentage cannot be traced to a summary row and submission artifacts.
- **Unverified update:** a post-submission result is presented as MLCommons-verified.
- **TDP energy claim:** board power replaces metered full-system energy.
- **Invented economics:** price, utilization, tariff, or duty cycle lacks dated buyer evidence.
- **Benchmark equals SLA:** passing public accuracy and latency constraints is treated as proof for a different production trace.

Conclusion

MLPerf Inference 6.1 is a rich procurement evidence set, but its value depends on disciplined reading. The release adds RAG, agentic inference, a more production-like API harness, new optimizations, and new available and preview hardware. Those additions broaden coverage while making cohort construction more important.

The defensible workflow begins with workload fit, filters like with like, reads the complete system and software configuration, and labels every normalization as descriptive. The clean four-to-16 accelerator example shows how scale efficiency can be calculated without pretending it predicts another deployment. The lack of power results for the sampled rows shows equal discipline: leave energy blank until it is measured.

For an enterprise buyer, the public benchmark should end where internal evidence begins. Freeze the exact artifacts, reproduce accuracy and performance, replay the intended traffic, meter power, validate facility constraints, and attach dated commercial quotes. Only then can benchmark output become an RFP score, a

price-performance ratio, or an investment case.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://mlcommons.org/2026/09/chairs-mlperf-inference-v6-1/
2. https://mlcommons.org/2026/09/mlperf-inference-v6-1-results/
3. https://github.com/mlcommons/inference_policies/blob/master/inference_rules.adoc
4. https://raw.githubusercontent.com/mlcommons/inference_results_v6.1/main/summary.csv
5. https://gpusmith.com/
6. https://docs.mlcommons.org/inference/submission/
7. https://github.com/mlcommons/inference_results_v6.1
8. https://raw.githubusercontent.com/mlcommons/inference_policies/master/inference_rules.adoc
9. https://mlcommons.org/2026/08/endtoend-inference/
10. https://github.com/mlcommons/inference/blob/master/tools/submission/submission_structure.md
11. https://gpusmith.com/articles/en/amd-mi350p-retrofit-guide
12. https://lambda.ai/blog/mlperf-inference-v6.1
13. https://raw.githubusercontent.com/mlcommons/policies/master/submission_rules.adoc
14. https://www.amd.com/en/products/accelerators/instinct/mi350/mi350p.html
15. https://www.amd.com/en/blogs/2026/amd-delivers-its-broadest-mlperf-inference-6-1-submission.html
16. https://gpusmith.com/articles/en/gpu-data-center-cooling-design
17. https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/components.html
18. https://gpusmith.com/articles/en/h200-vs-b200-vs-b300-vs-mi325x
19. https://www.dell.com/support/manuals/en-tm/poweredge-xe9785/xe9785_ism/poweredge-xe9785-system-overview?guid=guid-bd619e1a-540d-4a25-ada4-94f21a27295f&lang=en-us
20. https://www.supermicro.com/fr-fr/products/system/gpu/4u/as-4126gs-nmr-lcc
21. https://www.hpe.com/us/en/collaterals/collateral.c04447905.html
22. https://www.intel.com/content/www/us/en/newsroom/news/client-computing/intel-core-ultra-series-3-with-vpro-powers-next-gen-pcs-on-18a.html
23. https://infohub.delltechnologies.com/en-us/p/driving-ai-innovation-dell-technologies-advances-ai-inference-performance-with-mlperf-inference-v6-1/
24. https://www.intel.com/content/www/us/en/newsroom/news/data-center/intel-software-optimizations-boost-ai-inference-in-mlperf-v6-1.html
25. https://github.com/mlcommons/policies/blob/master/submission_rules.adoc
26. https://rocm.blogs.amd.com/artificial-intelligence/mlperf-inf-v6.1/README.html
27. https://blogs.nvidia.com/blog/vera-rubin-nvl72-mlperf-inference/
28. https://gpusmith.com/articles/en/gpu-rack-power-requirements-planning-guide
29. https://gpusmith.com/articles/en/gpu-cluster-acceptance-testing-runbook

30. https://www.nist.gov/itl/ai/software-performance-measurement
31. https://ftp.spec.org/cpu2026/docs/runrules.html
32. https://raw.githubusercontent.com/mlcommons/inference_results_v6.1/main/closed/AMD/src/gpt-oss-120b/offline_mi355x.yaml
33. https://raw.githubusercontent.com/mlcommons/inference_results_v6.1/main/closed/Crusoe/systems/512xMI355X_Crusoe_GPT.json
34. https://raw.githubusercontent.com/mlcommons/inference_results_v6.1/main/closed/Crusoe/systems/GB200-NVL72_GB200-186GB_aarch64x16_Dynamo.json
35. https://raw.githubusercontent.com/mlcommons/inference_results_v6.1/main/closed/Dell_MangoBoost/measurements/XE9785L/gpt-oss-120b/Interactive/user.conf
36. https://github.com/mlcommons/inference_policies/blob/master/power_measurement.adoc
37. https://www.energy.gov/cmei/femp/articles/take-five-life-cycle-costs-acquisition
38. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report_1.pdf?_bhlid=6e3cafbceb951012707f7cea7d3bd8f1bfdde416
39. https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/cooling.html
40. https://coreweave.com/blog/coreweave-leads-cloud-providers-in-mlperf-r-inference-v6-1-performance-with-nvidia-blackwell-ultra
41. https://gpusmith.com/articles/en/infiniband-vs-spectrum-x-vs-roce-vs-ethernet-ai-clusters

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com