

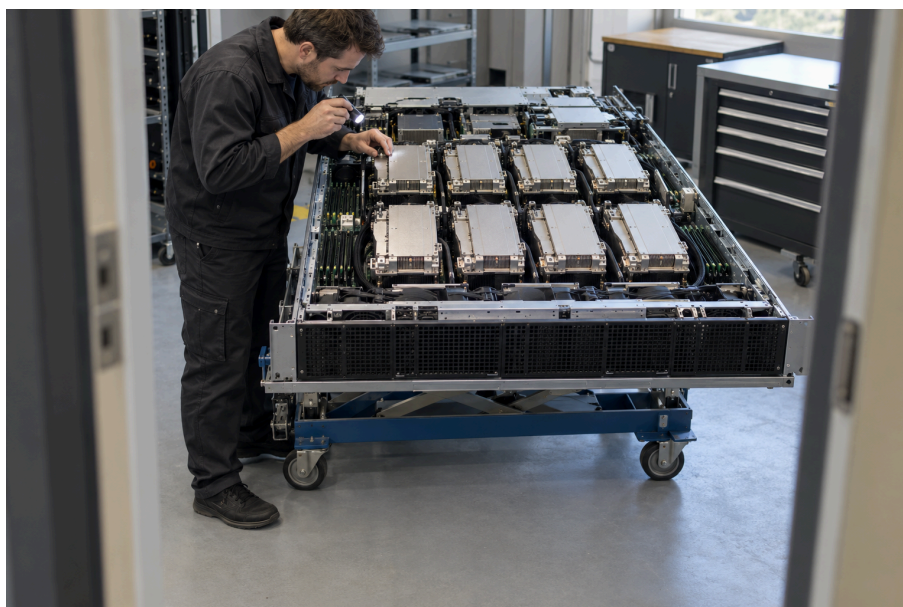


MLPerf Training 6.0 GPU Cost: Time to Quality Worksheet

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 · 2026-10-21 · mlperf training 6.0 gpu cost · mlperf training · gpu procurement · time to quality · gpu hours · training benchmarks · training cost · gpu comparison · reproducibility

Original article: <https://gpusmith.com/articles/en/mlperf-training-6-0-procurement-worksheet>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Key Changes
 4. Implementation Considerations and Process Changes
 5. Comparable Cohorts and the Procurement Evidence Ledger
 6. Reproducing a Public Submission
 7. GPU-Hour Cost and Facility Energy Worksheet
 8. Data Analysis and Evidence
 9. Implications and Future Directions
 10. Frequently Asked Questions (FAQs)
 11. Conclusion
-

Executive Summary

MLPerf Training 6.0 GPU cost cannot be read directly from a leaderboard. The benchmark measures wall-clock time to a specified dataset and quality target. (mlcommons.org) A buyer must combine a comparable result with a dated system or cloud quote, the full allocation window, and operating assumptions. MLCommons released the cohort on **June 16, 2026**. (mlcommons.org) The additions included **DeepSeek V3** and **GPT-OSS 20B**, both sparse-computation workloads. (mlcommons.org) Their inclusion expands workload coverage, rather than establishing a universal ranking for enterprise training.

A public submission supplies a useful, deliberately narrow example. A pinned Dell artifact describes a node containing **eight AMD Instinct MI355X accelerators**, each with **288 gigabytes (GB)** of memory. (github.com) One associated low-rank adaptation (**LoRA**) fine-tuning log records **487.78 seconds**, or **8.13 minutes**, and a final evaluation loss below its **0.925** target. Multiplying the accelerator count by that interval gives **1.084 GPU-hours**, rounded. (github.com) This is an analyst calculation for one recorded benchmark interval, excluding initialization, other runs, and allocation overhead. It is neither the official aggregated score nor a reproduced measurement from this report.

Public prices illustrate why the commercial boundary matters. The Amazon Web Services (**AWS**) Capacity Blocks table accessed for this report lists **p5.48xlarge at \$47.757 per instance-hour**. (aws.amazon.com) CoreWeave lists a North America **eight-GPU HGX H100 node at \$49.24 per hour on demand**. (coreweave.com) Those are different purchasing models. Neither rate prices the Dell run, which uses different accelerators and a different system. AWS also separates reservation and operating-system fees. (aws.amazon.com) A credible cost comparison therefore needs matched configurations and billing terms before multiplying time by price.

The procurement recommendation is to shortlist on relevant, valid time to quality, then authorize purchase only after an auditable pilot. Preserve the submission revision, system description, configuration, convergence logs, repeated-run spread, commercial quote, and measured energy boundary. Throughput charts remain useful screening evidence: Lambda explicitly measures samples processed per second, whereas the training

benchmark measures reaching quality. GPU telemetry also has a narrower boundary than whole-system or facility energy. Unknown acquisition prices, staffing effort, and facility consumption should remain explicit worksheet blanks until measured or quoted. (lambdai.com) (docs.nvidia.com)

Introduction and Background

An enterprise buyer of **graphics processing units (GPUs)** needs an acceptable trained model delivered within a practical budget and schedule. A faster accelerator can be relevant to that decision, but the procurement unit is usually a usable system: processors, memory, networking, storage, software, and access to capacity. The central question is which public result provides enough evidence to justify a paid pilot, followed by which pilot observations justify purchase.

MLPerf Training supplies a defined endpoint: training a model on a specified dataset until the required quality is achieved. (mlcommons.org) (www.digitalocean.com) That endpoint distinguishes useful model progress from processing examples quickly. It does not, by itself, provide a commercial quote or show how a buyer's private training workload will converge.

This report treats GPU Smith as an adjacent independent engineering adviser. Its perspective is an evidence standard for procurement, rather than an accelerator or cloud option in the comparison. The analysis covers model training, including a public fine-tuning example. It does not use inference serving performance to estimate training time.

The analysis is anchored to **October 21, 2026** access checks. Version control matters because the current benchmark landing page already includes workloads marked **v6.1**, alongside retained **v6.0** entries. The worked example therefore uses immutable submission links rather than treating the entire current page as the historical cohort. (mlcommons.org)

Readers should distinguish three kinds of statement throughout: a benchmark body's definition, a submitter's recorded result, and an analyst's procurement calculation. The selected submission was inspected, not executed again. No facility meter readings or buyer acquisition quotes were obtained. Those omissions limit the report's cost conclusion: it can demonstrate the worksheet and identify evidence requirements, but it cannot declare a purchased system's measured cost to train.

Key Changes

Sparse workloads make task selection more consequential

The release added **DeepSeek V3** and **GPT-OSS 20B**, extending coverage of mixture-of-experts (**MoE**) training. MoE models route tokens through selected experts, so a procurement test should examine the relevant sparse workload rather than assume a dense-model result transfers automatically. The release also describes the smaller new benchmark as a representative portion of end-to-end training. Its elapsed time should not be advertised as the cost or duration of complete production pretraining. (mlcommons.org)

The practical change is a broader choice of evidence. A buyer running sparse pretraining should seek a sparse submission with relevant scale and topology. A buyer adapting an existing model should begin with a fine-tuning task and a suitable quality threshold. Matching the task is the first filter; accelerator branding comes

later.

The result is a system and software observation

Closed division requires the reference preprocessing, model, training method, and quality target. **Open division** permits alternative data, preprocessing, models, or methods. Both can be informative, but their claims need different interpretation. The current official rules provide these definitions; buyers should retain the rules applicable to the original round rather than assume the latest document freezes every historical detail. (github.com)

This distinction prevents an overly broad claim that Closed makes every implementation identical. It establishes controlled problem definitions while still leaving system and implementation details important to a procurement decision. Open evidence can support an interesting alternative method, provided the report explains the changed problem or implementation.

Availability and validity belong in the shortlist

MLCommons distinguishes systems containing components available for purchase or cloud rental from preview configurations. Availability classification helps screen what can be ordered, but it does not replace a supplier's confirmation of delivery, capacity, or the exact configuration. (mlcommons.org)

Published results also require a status check. MLCommons maintains a change log for modifications after publication. The buyer should record when that check occurred and preserve the result used in the decision, rather than rely indefinitely on a launch announcement. (mlcommons.org)

A useful intake checklist is:

- **Task:** Specify pretraining, fine-tuning, or the exact training operation.
- **Endpoint:** State the dataset, evaluation procedure, and acceptable quality.
- **Division:** Preserve the declared comparison category.
- **Availability:** Verify the configuration can actually be supplied.
- **Revision:** Retain the submission and applicable rule revisions.
- **Status:** Check the result change log before issuing the request for proposal.

These are proposed procurement controls. Their purpose is to make the evidence package reconstructable when a reviewer asks why a particular result entered the shortlist.

Implementation Considerations and Process Changes

Separate throughput from time to quality

Throughput answers how many samples a system processes per second. Lambda explicitly includes tokens and images in that definition and notes that implementation affects throughput. **Time to quality** answers how long training takes to reach a specified evaluation target. Converting between them requires knowing the amount of training needed to converge, not merely the speed of an iteration. (lambda.ai) (mlcommons.org)

For initial screening, a throughput chart can identify plausible candidates. Before calculating ratios, however, normalize the model, batch size, numerical precision, optimizer, software, and communication arrangement. AIME explicitly notes that batch size affects both performance and training results. (www.aime.info) Its page also discloses distinct software environments, including **PyTorch 2.5.1, CUDA 12.1.105, and Python 3.10.12**. (www.aime.info) These details are comparison inputs, not footnotes to discard.

Lambda's page discloses a legacy environment containing **Ubuntu 20.04, PyTorch 1.13.0a0+d0d6b1f, CUDA 11.8.0, and cuDNN 8.6.0.163**. (lambda.ai) That disclosure makes the chart interpretable, but it does not turn its values into measurements of a newer submission stack. Avoid creating a GPU speedup ratio from charts that changed software and hardware together.

Read the workload label before using a chart

DigitalOcean's training guide points readers to MLPerf as evidence about training speed to target quality. A LessWrong discussion comment also links Lambda's standardized benchmarks in compute-estimation methodology. These references show how benchmark methods can be reused; they do not independently validate a particular GPU ranking. (www.digitalocean.com) (www.lesswrong.com)

A fetched Runpod comparison page labels its language-model measurements as **September 2026 vLLM** benchmarks. That is serving evidence rather than training convergence evidence. Procurement teams should classify the measured operation before using any attractive chart in a training business case. (www.runpod.io)

The operational comparison checklist is:

- **Model:** Match architecture and model size.
- **Data:** Match preprocessing and sequence or sample shape.
- **Batch:** Record local batch, global batch, and accumulation.
- **Precision:** Preserve numerical formats and scaling recipes.
- **Optimizer:** Preserve settings that affect convergence.
- **Software:** Record framework, libraries, compiler, and container.
- **Network:** Record fabric, interface selection, and parallelism.
- **Metric:** Keep throughput and time-to-quality ratios distinct.

Do not combine a training numerator with an inference denominator. A chart may still guide a separate capacity question, but a training cost worksheet should contain evidence for the training operation being purchased.

Comparable Cohorts and the Procurement Evidence Ledger

Decide what comparison the evidence permits

A **comparable cohort** is a buyer-defined set of results sufficiently similar to support a particular decision. It should be narrower than the full release. Closed division is a useful condition, but procurement also needs workload relevance, system availability, and software that the buyer can obtain. The official division definitions and availability categories address different parts of this problem.

Table 1 below is a proposed comparable-cohort gate. It identifies what to retain before interpreting a ranking.

Field	Evidence to retain	Procurement interpretation
Release and status	Historical submission revision and current change-log check. (mlcommons.org)	Compare a declared cohort with traceable status.
Workload	Model, dataset, preprocessing, and quality target.	A similar model name alone is insufficient.
Division	Closed or Open declaration and applicable rules.	Explain any permitted methodological differences.
Availability	Available or preview classification, plus supplier confirmation.	Separate public classification from an orderable quote.
Topology	Node count, accelerators, host processors, memory, and fabric.	Preserve the full system being compared.
Software	Submitted environment and workload configuration.	Treat stack changes as explicit experimental changes.
Runs	Individual logs and the applicable aggregation method. (github.com)	Do not promote one selected run into an official score.
Commercial scope	Dated capacity or acquisition quote and exclusions. (aws.amazon.com)	Align the billed unit with the system and duration.

The table is an evidence filter, not a scorecard assigning universal weights. A mismatch should trigger a documented explanation or a pilot designed to resolve it. The outcome can be a conditional shortlist rather than a forced ranking of systems that solved different problems.

Preserve enough detail to reproduce the decision

The ledger should separate public evidence from buyer observations. The public record may establish a submitted configuration and a successful benchmark interval. The pilot must establish whether the quoted configuration, buyer data path, and operating procedures deliver acceptable results. Keep those two records linked without merging their provenance.

Recommended ledger fields are:

- **Identity:** URL, repository revision, path, and retrieval date.
- **Integrity:** Hashes for downloaded code, data manifests, and containers.
- **Configuration:** Launch arguments, environment, and power settings.
- **Convergence:** Evaluation sequence, target crossing, and run completion.
- **Distribution:** All required runs and the scoring calculation.
- **Economics:** Quote identifier, billing window, storage, and labour.
- **Energy:** Meter location, units, sampling interval, and timestamps.

These fields are this report's proposed procurement method. They should not be read as a claim that every public submission already supplies buyer-generated hashes, facility measurements, or prices.

Reproducing a Public Submission

What the inspected artifacts actually establish

The worked example uses Dell's versioned **1xE9785L_8xMI355X** system record and its **llama2_70b_lora/result_6.txt** log. The system record identifies **one node with eight MI355X accelerators** and classifies it as Closed and Available on-premise. It records **Ubuntu 24.04 LTS and ROCm 7.0.1**. These are attributes of the submitted system, not a guarantee about every system carrying the same accelerator name. ([github.com](#))

Table 2 below records the evidence boundary for the selected run.

Item	Observed public evidence	Limit on interpretation
System	Eight AMD Instinct MI355X accelerators, 288 GB of high-bandwidth memory each. (github.com)	The complete configured node is the procurement reference.
Workload	Llama 2 70B LoRA, with target evaluation loss 0.925. (github.com)	This is fine-tuning evidence, not full-model pretraining cost.
Dataset	The versioned-cohort reference table specifies SCROLLS GovReport. (github.com)	Retain the exact dataset preparation and evaluation procedure.
Individual duration	487.78 seconds, displayed as 8.13 minutes. (github.com)	One recorded run, excluding other allocation intervals.
Endpoint	Final evaluation value 0.9230323433876038 and successful completion. (github.com)	Benchmark quality does not establish buyer-specific model utility.
Accelerator-time calculation	Multiply the system's accelerator count by the recorded interval in hours.	Analyst arithmetic, not an official aggregate or invoice.
Missing cost inputs	Quoted matching system rate, setup time, staffing, and energy.	No measured dollar cost or facility energy is asserted.

This example demonstrates traceability rather than a vendor recommendation. The retrieved official rules describe discarding the fastest and slowest runs and averaging the remainder. The selected individual log cannot substitute for the cohort's aggregate result, and no aggregate for this selected submission was extracted here. ([github.com](#))

Reproduction procedure for a paid pilot

AMD's vendor-specific reproduction guide covers environment setup, data preparation, configuration, execution, and validation. ([rocm.blogs.amd.com](#)) Its LoRA preparation notes **270 GB** of disk space for model download and conversion. ([rocm.blogs.amd.com](#)) It also documents setting **NEXP to 10** for a multi-run result. ([rocm.blogs.amd.com](#)) Those details illustrate why preparation and repetition belong in the pilot budget. The guide does not establish transferability to another vendor or configuration.

Start from the exact submission, then apply a controlled procedure:

- **Freeze artifacts:** Resolve revisions and container digests before scheduling capacity.

- **Prepare inputs:** Record dataset manifests, checkpoint provenance, and local hashes.
- **Verify hardware:** Compare the delivered node and fabric with the evidence ledger.
- **Execute unchanged:** Attempt the submitted configuration before modifying it.
- **Retain every run:** Preserve successes, incomplete runs, and operational delays.
- **Recompute scoring:** Apply the correct historical run count and aggregation rule.
- **Repeat buyer workload:** Evaluate the model task that will authorize purchase.

PyTorch warns that complete reproducibility is not guaranteed across releases, commits, or platforms. It also notes that deterministic operations can be slower. Accordingly, a pilot should report distribution and environment, and should not silently change determinism settings to improve either repeatability or speed. (docs.pytorch.org)

For pilot analysis, report every elapsed time and quality outcome, then provide the median, minimum, maximum, and sample standard deviation alongside the required benchmark aggregate. Keep the complete allocation history even when the official scoring procedure excludes an extreme run. This is a proposed operational report, rather than an additional official scoring requirement.

GPU-Hour Cost and Facility Energy Worksheet

Choose the billed unit before multiplying

For a fixed allocation, **GPU-hours = allocated accelerators × allocation hours**. For changing allocations, sum that product over each interval. Benchmark GPU-hours use the benchmark interval; billed GPU-hours use the reservation or allocation interval. The public run yields the benchmark-interval quantity calculated below, rather than a complete pilot bill.

Use **node-hours × quoted node-hour rate** when the supplier sells nodes. Do not add a separate GPU charge when the machine rate already includes it: Google's accelerator-optimized machine pricing explicitly includes GPU cost. Conversely, Google's reservation documentation says reserved resources can be billed even when not started. Billing and useful training time can therefore diverge. (cloud.google.com)

Table 3 below is the **MLPerf Training 6.0 procurement worksheet**. Here G is accelerator count, N is node count, T_bench is benchmark seconds, H_alloc is allocated hours, and R_node is currency per node-hour. Information technology (IT) energy uses the declared equipment boundary. All unobserved variables remain inputs to obtain, rather than estimated benchmark facts.

Cost component	Calculation and units	Required evidence or sensitivity
Benchmark accelerator time	$G \times T_{\text{bench}} / 3,600$, GPU-hours when T_bench is seconds.	Use the public count and recorded duration from Table 2.
Cloud allocation	$N \times H_{\text{alloc}} \times R_{\text{node}}$, currency.	Quoted nodes, reserved hours, rate, region, and terms.
Setup and repetitions	Sum all charged allocation intervals.	Include data staging, validation, repeated runs, and recovery.
Storage and transfer	Itemized quantities × quoted unit prices.	Azure notes virtual machine (VM) pricing includes local disk only. (azure.microsoft.com)

Cost component	Calculation and units	Required evidence or sensitivity
Measured IT energy	Sum input kW × interval hours, kWh.	Include the defined compute, storage, and network boundary. (www1.eere.energy.gov)
Facility allocation	IT kWh × applicable overhead factor.	The Department of Energy (DOE) defines power usage effectiveness (PUE) using facility and IT energy. (www.energy.gov)
Electricity	Facility kWh × applicable tariff, currency.	Local tariff, time periods, and separate demand charges where applicable. (www.eia.gov)
Capital rate	Configured purchase price / productive lifetime system-hours, currency per system-hour.	Quote the full system; vary lifetime and productive utilization explicitly.
Allocated pilot capital cost	Capital rate × pilot system-hours, currency.	Count pilot system-hours on the same productive-utilization basis as the lifetime denominator.
Operations	Staff hours × loaded labour rate, plus allocated support.	Buyer estimates and supplier maintenance terms.
Total allocated pilot cost	(Cloud allocation or allocated pilot capital cost) + nonincluded operating costs, currency.	Document inclusion boundaries to prevent double counting; report cash paid for equipment separately from allocated pilot cost.

This worksheet supports a cost-to-quality decision only when the corresponding pilot reaches quality. A low hourly rate is insufficient if the allocation takes longer, requires more repetitions, or includes idle reservations. Conversely, faster time may justify a premium when a training deadline has business value. That value should be stated separately from infrastructure cost.

Measure energy at the intended boundary

Power usage effectiveness (PUE) is the ratio of total annual facility energy to annual information technology (IT) energy. Applying a facility overhead factor to a short training run is an allocation assumption, not a direct facility measurement. If facility energy is measured already, do not multiply it by PUE again. (www.energy.gov)

Device telemetry can help diagnose utilization and integrate GPU energy. NVIDIA's **Data Center GPU Manager (DCGM)** aggregates workload statistics across participating GPUs and requires recording before the workload starts. Its GPU boundary does not establish total server, shared storage, network, or cooling energy. (docs.nvidia.com) (www1.eere.energy.gov)

Data Analysis and Evidence

Public prices provide inputs, not a cost ranking

On the access date, the AWS Capacity Blocks table listed **eight-H100 p5.48xlarge at \$47.757 per instance-hour**, including a US East Atlanta Local Zone entry. Its displayed accelerator-hour equivalent was **\$5.970**, rounded. This is reservation pricing, with a separate operating-system fee, rather than an interchangeable on-demand training quote. (aws.amazon.com)

CoreWeave's North America page listed **eight-GPU HGX H100 at \$49.24 per hour** and **HGX H200 at \$50.44 per hour**, both on demand. It also states that its individual-GPU pricing applies exclusively to inference-platform customers. A training buyer should therefore use the applicable node price and obtain a dated quote for the required capacity. (coreweave.com)

These are price-only observations. No matched time-to-quality result is paired with either price in this report. They cannot establish which provider delivers cheaper training, and neither may be multiplied by the Dell benchmark's duration to claim a measured cost.

Other opened sources reveal useful missing inputs. Oracle lists an **eight-MI355X shape with 288 GB per accelerator**, but an attributable hourly price was not recovered. A matching accelerator label would still leave topology and software to verify. Google documents region and zone restrictions; Azure separates managed storage from its VM price. These facts help shape quote requests without filling unknown cells with third-party estimates. (www.oracle.com) (cloud.google.com) (azure.microsoft.com)

Quantify what is known and leave the rest conditional

For the selected public log, the calculation is $8 \times 487.78 / 3,600 = 1.0839556$ GPU-hours before rounding. (github.com) If a matching deployment had a genuinely applicable quoted accelerator-hour rate r , its benchmark-interval charge would be $1.0839556 \times r$. No value for r has been established for that submitted configuration. A node-based quote instead needs the corresponding node duration and actual billing window.

Historical electricity averages illustrate a different uncertainty. The U.S. Energy Information Administration (EIA) reports **2025 industrial and commercial averages of 8.62 and 13.41 cents per kWh**, respectively, in its page citing preliminary data from **February 2026**. (www.eia.gov) They are contextual statistics, not the buyer's tariff. Locality and supply conditions affect electricity prices, so the worksheet should use the applicable facility schedule and contractual charges. (www.eia.gov)

Power configuration also needs verification. Dell documents an optional permanent cap of **1 kW per GPU** for specified XE9785 configurations. That is a configuration capability for the documented system, not a measured draw or proof that the selected XE9785L submission used that cap. Ask explicitly for the delivered system's power settings. (www.dell.com)

The sensitivity report should vary allocation duration, quoted rate, productive utilization, and measured input power independently. Establish low, base, and high inputs from pilot observations and dated quote alternatives. For fixed accelerator count G and time bounds T_{low} and T_{high} in hours, the benchmark range is $G \times T_{\text{low}}$ through $G \times T_{\text{high}}$ GPU-hours. Report the source of each bound; leave the range unpopulated when evidence is missing. Do not describe an assumed range as a statistical confidence interval.

Implications and Future Directions

The main procurement change is procedural: treat a benchmark result as an auditable experimental record, then connect it to a commercial configuration and the buyer's workload. The evidence in Table 2 makes a narrow accelerator-time calculation possible; it leaves the purchase quote and operational evidence to be obtained.

A network-sensitive deployment deserves a network test before a paid training campaign. NVIDIA documents that its Collective Communications Library (**NCCL**) automatically detects interfaces for inter-node communication. The pilot should preserve interface selection and communication settings alongside the topology, rather than assume a successful container launch verifies the fabric. This advice applies to the relevant software stack, not automatically to every vendor's communication libraries. (docs.nvidia.com)

Checkpoint behavior also matters operationally. PyTorch's distributed checkpoint documentation describes resharding across different cluster topologies. That capability can support recovery planning, but the buyer must test the actual implementation and record the time and storage it consumes. It should not be treated as evidence that the selected public benchmark measured production recovery. (docs.pytorch.org)

For an independent adviser such as GPU Smith, the appropriate purchase gate is a documented acceptance decision. The recommendation is to require:

- **Relevance:** A declared buyer workload and meaningful quality endpoint.
- **Traceability:** Public submission artifacts linked to the tested configuration.
- **Repeatability:** An explained run distribution and preserved convergence evidence.
- **Economics:** A dated quote joined to measured allocation and operating inputs.
- **Operations:** Demonstrated data staging, checkpointing, and recovery procedures.
- **Facilities:** Infrastructure and facilities engineering approval of the measured boundary.

Approval should follow the buyer's predeclared limits for quality, elapsed time, cost, and delivery. If a public result satisfies only the first screening stage, the next authorized expenditure is a bounded pilot. The unresolved evidence should be named in the request for proposal, rather than converted into a generalized promise about accelerator performance.

Frequently Asked Questions (FAQs)

What is MLPerf Training 6.0 time to quality?

It is the elapsed benchmark time needed to reach the specified quality target on the specified dataset. It is not synonymous with a full production training campaign: the release describes a representative training portion for the smaller new sparse workload. Keep the benchmark interval and the buyer's complete job separate. (mlcommons.org)

How is MLPerf Training 6.0 GPU-hour cost calculated?

Multiply the allocated accelerator count by elapsed hours, then apply a rate only when its system, billing unit, and commercial terms match. Use the observed interval for benchmark accelerator-time, then add separately measured allocation overhead before pricing. The worksheet demonstrates that calculation without assigning an unverified matching rate.

How should an MLPerf Training 6.0 GPU comparison guide procurement?

Use a relevant division, workload, system description, and availability category to form a shortlist. Then reproduce the selected configuration and test the buyer workload under quoted conditions. Open and Closed results can both be useful, but their permitted methodological differences must remain visible.

(mlcommons.org)

Can a vendor reproduction guide prove another system's cost to train?

A guide provides a procedure for its documented environment. It does not establish another system's convergence, delivered configuration, or commercial price. AMD's reproduction documentation is useful engineering evidence for its own submissions; cross-system conclusions require the corresponding artifacts and pilot observations. (rocm.blogs.amd.com)

Conclusion

MLPerf Training can support procurement when the buyer preserves the experiment it describes. The relevant evidence is a declared workload reaching a declared quality target on a documented system. A throughput chart, an accelerator specification, and an hourly price each answer part of the decision, but they do not automatically combine into a validated training-cost claim.

The public example shows the useful boundary. Its individual duration and accelerator count support a transparent GPU-hour calculation. They do not establish the official aggregate score, the allocation needed for a complete pilot, or the price of a matching deployment. Keeping those distinctions explicit makes the worksheet reusable without implying a purchase recommendation that the evidence does not support.

A disciplined paid pilot closes the remaining gaps. It preserves revisions and hashes, reproduces the submitted configuration, records convergence across runs, measures the appropriate energy boundary, and joins the resulting allocation to a dated quote. It also tests the buyer's data pipeline and operational requirements, because benchmark completion alone does not authorize every production use.

Purchase should follow evidence that satisfies predeclared quality, schedule, cost, and delivery criteria. Until then, unknown prices and operating inputs should remain visible. The strongest shortlist is the one whose unresolved questions can be converted into specific pilot measurements, and whose eventual cost-to-quality conclusion another engineer can reconstruct.

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/benchmarks/training>
2. <https://mlcommons.org/2026/06/mlperf-training-v6-0-results>
3. https://github.com/mlcommons/training_results_v6.0/blob/eabf23a07b2a0c60a289ff871dc3a46fff0d0421/Dell/systems/1xXE9785L_8xMI355X.json
4. https://github.com/mlcommons/training_results_v6.0/blob/eabf23a07b2a0c60a289ff871dc3a46fff0d0421/Dell/results/1xXE9785L_8xMI355X/llama2_70b_lora/result_6.txt

5. <https://aws.amazon.com/ec2/capacityblocks/pricing>
6. <https://coreweave.com/pricing>
7. <https://lambda.ai/gpu-benchmarks>
8. <https://docs.nvidia.com/datacenter/dcgm/latest/user-guide/feature-overview.html>
9. <https://www.digitalocean.com/solutions/gpu-model-training>
10. https://github.com/mlcommons/training_policies/blob/master/training_rules.adoc
11. <https://mlcommons.org/results-change-log>
12. <https://www.aime.info/blog/en/pytorch-2-gpu-performace-benchmark-comparison>
13. <https://www.lesswrong.com/posts/HvqQm6o8KnwxbdmhZ/estimating-training-compute-of-deep-learning-models>
14. https://www.runpod.io/gpu-compare/rtx-3090-vs-h100-pcie?98063dc5_page=2
15. <https://github.com/mlcommons/training>
16. <https://rocm.blogs.amd.com/artificial-intelligence/mlperf-training6.0-repro/README.html>
17. <https://docs.pytorch.org/docs/2.14/notes/randomness.html>
18. <https://cloud.google.com/products/compute/gpus-pricing?hl=en>
19. <https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux>
20. https://www1.eere.energy.gov/femp/pdfs/hpc_metering_protocol.pdf
21. <https://www.energy.gov/cmei/femp/cooling-water-efficiency-opportunities-federal-data-centers>
22. <https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php>
23. <https://www.oracle.com/cloud/price-list?isappinstalled=0>
24. https://www.dell.com/support/manuals/en-in/poweredge-xe9785/xe9785_ism/power-capping-configuration-with-mi355x-gpu?guid=guid-73e7b829-65b2-47bb-9c88-9b73659eb35e&lang=en-us
25. https://docs.nvidia.com/deeplearning/ncccl/user-guide/docs/troubleshooting/networking_troubleshooting.html
26. https://docs.pytorch.org/tutorials/recipes/distributed_checkpoint_recipe.html

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>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.