PULSE BENCHMARKS

GPU Compute Pricing Assessments

Methodology Specification

Version 1.0

April 2026

Pulse Data Co.

1. Purpose and Scope

The Pulse Benchmarks is a family of daily price assessments for cloud GPU compute. Each assessment reports the cost of renting one GPU for one hour from a defined segment of the cloud compute market. The methodology follows Price Reporting Agency (PRA) principles: transparent data collection, documented normalization rules, source-family segmentation, and auditable provenance.

This document is the governing methodology specification for all Pulse benchmark series. Any change to the rules described here constitutes a methodology revision and must be versioned, logged in the Data Changelog, and communicated to subscribers before taking effect.

1.1 Benchmark Series Covered

Version 1.0 of this methodology governs four benchmark series:

Series Name	GPU Model	Source Family	Pricing Type
Pulse H100 SXM Hyperscaler On-Demand	NVIDIA H100 SXM 80GB	Hyperscaler	On-Demand
Pulse H100 SXM Neocloud On-Demand	NVIDIA H100 SXM 80GB	Neocloud	On-Demand
Pulse A100 80GB Hyperscaler On-Demand	NVIDIA A100 80GB	Hyperscaler	On-Demand
Pulse A100 80GB Neocloud On-Demand	NVIDIA A100 80GB	Neocloud	On-Demand

Additional series (spot pricing, marketplace assessments, additional GPU models) may be added in future methodology versions as data coverage meets the publishability thresholds defined in Section 6.

2. Key Definitions

Assessment: A single evaluated price point for one GPU model, from one provider, for one pricing type, on one assessment date. The atomic unit of the Pulse dataset.

Benchmark Series: A named, published time series that aggregates assessments across providers within a single source family. For example, "Pulse H100 SXM Hyperscaler On-Demand" is one benchmark series.

Source Family: A classification of cloud compute providers by market segment. Pulse defines three source families: Hyperscaler, Neocloud, and Marketplace. Providers within a family are more comparable to each other than to providers in other families, in terms of reliability, procurement experience, and buyer expectations. Source families are never blended into a single headline number.

Hyperscaler: A large-scale public cloud platform with global infrastructure, enterprise SLAs, and integrated service ecosystems. Currently: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and Oracle Cloud Infrastructure (OCI).

Neocloud: A GPU-focused cloud provider offering dedicated compute instances at published list prices. Neoclouds typically offer simpler procurement (no reserved instances or committed-use contracts as the primary model) and narrower service scope than hyperscalers. Currently: Lambda Labs, RunPod, CoreWeave, Paperspace, and DataCrunch.

Marketplace: A platform where independent suppliers list GPU capacity at individually set prices. The assessment for a marketplace reflects aggregate listing statistics (median, percentiles) rather than a single provider's list price. Currently: Vast.ai. Marketplace assessments are governed by separate publishability criteria and are not included in Version 1.0.

On-Demand Pricing: The standard hourly rate for GPU compute available without advance commitment, reservation, or bidding. On-demand prices reflect what a buyer would pay to provision a GPU immediately at the provider's published rate.

Assessed Price (USD/GPU-hour): The normalized, all-in hourly cost of renting one GPU from a given provider. For providers that price per-instance (bundling CPU, RAM, storage, networking, and GPU into a single SKU), the assessed price is the full instance price divided by the number of GPUs. This is the primary Pulse benchmark figure. It represents what a buyer actually pays per GPU-hour and is directly reproducible from the provider's published price list. It does not attempt to decompose the instance price into GPU-only and non-GPU components. Where a provider exposes component pricing (e.g., GCP separates accelerator and VM compute costs), the all-in assessed price includes all components. A supplementary estimated accelerator-only figure may be published separately for hyperscaler series in future versions, clearly labelled as either directly observed (GCP) or modelled from comparable non-GPU instance pricing (AWS, Azure). Estimated component splits are supplementary analytics, not the primary benchmark.

Assessment Time: The fixed daily timestamp at which all Pulse assessments are anchored: 18:00 UTC. Data collected within the assessment window (typically 17:00–19:00 UTC) is attributed to this timestamp.

3. Source Family Segmentation

Pulse segments the cloud GPU market into three source families. This segmentation is the most important structural decision in the methodology. Each family represents a distinct product category with different reliability characteristics, procurement models, support levels, and buyer expectations. Blending prices across families would produce a number that is statistically calculable but commercially meaningless.

The price spread between families is structural, not noise. Hyperscaler pricing for a given GPU consistently runs several multiples above neocloud pricing for the same hardware. This spread reflects real differences in the product being purchased—infrastructure reliability, compliance, networking, support, ecosystem—not market inefficiency.

3.1 Family Definitions

Source Family	Providers (v1.0)	Product Characteristics
Hyperscaler	AWS, Azure, GCP, OCI	Global infrastructure, enterprise SLAs, compliance certifications, integ
Neocloud	Lambda, RunPod, CoreWeave, Hyperscale, DataCrunch	GPU-focused, limited procurement, published list pricing, narrower se
Marketplace	Vast.ai	Supply-side pricing by independent operators, variable reliability, lowe

3.2 Why Separate Benchmarks

The decision to publish separate benchmark series per source family (rather than one blended number per GPU) follows established PRA practice. Price reporting agencies routinely publish separate assessments for the same physical commodity when the delivery mechanism, contractual terms, or market structure differ materially. A CIF cargo price and an FOB price for the same crude are published separately because they represent different commercial transactions—even though the underlying molecule is identical.

The same logic applies to GPU compute. An H100 SXM hour on AWS and an H100 SXM hour on RunPod deliver the same peak FLOPS, but the product bundle (infrastructure reliability, networking, compliance, support, ecosystem) differs enough that a buyer choosing between them is making a segment decision, not a price-arbitrage decision. The Pulse methodology respects this by never blending across families.

4. Data Collection

4.1 Collection Schedule

All providers are collected daily. The assessment anchor time is 18:00 UTC. All providers are collected sequentially in a single automated run. Data from each run is attributed to the 18:00 UTC assessed_at timestamp for that date regardless of the exact clock time of collection.

Collection runs 365 days per year with no business-day distinction. GPU compute markets operate continuously and Pulse assessments reflect this.

4.2 Data Sources by Provider

4.2.1 Hyperscaler Sources

Amazon Web Services (AWS)

- Source: AWS Pricing API
- Pricing model: Per-instance-hour. Instance price divided by GPU count from instance_gpu_map reference table.

• Region: US East (N. Virginia) — us-east-1

• Instance types: p5.48xlarge (H100 SXM ×8), p4de.24xlarge (A100 80GB ×8)

• SKU filtering: AWS publishes dozens of pricing SKUs per instance type, varying by operating system, tenancy, capacity status, market option, and pre-installed software. Each combination carries a different price. The correct on-demand price is isolated by filtering to exactly one SKU per instance type: operatingSystem = "Linux", tenancy = "Shared", capacitystatus = "Used", marketoption = "OnDemand", and preInstalledSw = "NA". Without all five filters, non-GPU pricing components (Capacity Block reservation fees, OS license surcharges, SQL Server license bundles, dedicated tenancy premiums) are included as if they were standalone GPU prices. These filters are implemented in the AWS collector and constitute a normalization rule: any change to them changes assessed output.

Microsoft Azure

• Source: Azure Retail Prices REST API

• Pricing model: Per-VM-hour. VM price divided by GPU count from instance_gpu_map.

• Region: US East — eastus

• VM series: ND96asr_v4 / ND96amsr_A100_v4 (A100 80GB), ND96isr_H100_v5 (H100 SXM)

Google Cloud Platform (GCP)

• Source: Cloud Billing Catalog API v1

• Pricing model: Component pricing. GCP bills GPU accelerators separately from VM compute (CPU + RAM). For GPUs with a standard bundled VM family (A100 via A2, H100 via A3), the all-in assessed price is: standalone GPU rate + (VM CPU+RAM cost ÷ GPU count). Where a bundled VM path exists, the standalone accelerator rate alone is not separately published as a benchmark assessment, to avoid mixing different pricing instruments. For GPUs without a bundled VM family, the standalone GPU accelerator rate is used.

• Region: GCP publishes both Americas-level aggregate pricing and city-specific pricing. Pulse uses the Americas aggregate SKU where available. This effectively represents US-region pricing (us-central1, us-east1) without double-counting city-specific variants.

Oracle Cloud Infrastructure (OCI)

• Source: Oracle Cloud Pricing API (public, no credentials required)

• Pricing model: Per-GPU-hour (metricName = "GPU Per Hour"). Prices are already per-GPU; no instance normalization needed.

• Region: Uniform global pricing (same price in every OCI region).

• Shapes: BM.GPU.H100.8 (H100 SXM ×8), BM.GPU.A100-v2.8 (A100 80GB ×8). Bare metal only for flagship GPUs.

4.2.2 Neocloud Sources

Lambda Labs

• Source: REST API

- Pricing model: Direct per-GPU-hour

RunPod

- Source: GraphQL API (programmatic)
- Pricing model: Direct per-GPU-hour for single-GPU; per-instance normalized to per-GPU for multi-GPU

CoreWeave

- Source: HTML pricing page scrape
- Pricing model: Direct per-GPU-hour

Paperspace

- Source: HTML pricing page scrape
- Pricing model: Direct per-GPU-hour (single-GPU instances only; multi-GPU and committed-use pricing excluded)

DataCrunch (Verda Cloud)

- Source: REST API
- Pricing model: Per-instance normalized to per-GPU-hour. Smallest available instance (fewest GPUs) preferred as representative price.

4.3 GPU Model Canonicalization

Each provider uses different names and identifiers for the same physical GPU. Pulse maintains a canonical GPU model registry that maps all provider-specific names to a single canonical identifier. For example, AWS's "p5.48xlarge", Lambda's "gpu_1x_h100_sxm5", and Azure's "NVIDIA H100" all map to the canonical model ID "h100_sxm".

The mapping tables are maintained in collector source code and the instance_gpu_map database table. Any change to these mappings constitutes a methodology change and must be versioned accordingly.

4.4 Price Normalization

All assessed prices are expressed in USD per GPU-hour. The normalization method depends on the provider's pricing model:

Pricing Model	Normalization Rule	Providers
Per-GPU-hour	No normalization needed. Price used directly.	Lambda, RunPod, CoreWeave, Paperspace
Per-instance-hour	Instance price ÷ GPU count (from instance_gpu_map table)	AWS, Azure
Component pricing	Standalone GPU rate + (VM CPU+RAM cost ÷ GPU count)	GCP

Pricing Model	Normalization Rule	Providers
Per-instance (API)	Instance price ÷ GPU count; smallest instance preferred	DataCrunch

For providers that offer multiple instance sizes with the same GPU (e.g., GCP's A2 family offers 1, 2, 4, and 8 A100 GPU configurations), the default representative configuration is the smallest standard instance, unless a different configuration is designated in the benchmark specification for comparability reasons. The smallest instance is the default because it most closely reflects per-GPU economics comparable to providers that offer single-GPU instances.

4.5 Bundling Disclosure

Hyperscaler GPU instances bundle non-GPU resources (vCPUs, RAM, local storage, and networking) into a single instance price. When the Pulse assessed price divides instance price by GPU count, the resulting per-GPU-hour figure includes the cost of these bundled resources. This is a known and intentional simplification. The following table documents the bundle composition of each assessed hyperscaler instance as of April 2026:

Instance	GPU	GPUs	vCPUs	RAM	Storage	Network
p5.48xlarge	H100 SXM	8	192	2,048 GB	8x 3.84 TB	3,200 Gbps EFA
p4de.24xlarge	A100 80GB	8	96	1,152 GB	8x 1 TB	400 Gbps EFA
p4d.24xlarge	A100 40GB	8	96	1,152 GB	8x 1 TB	400 Gbps EFA
ND96isr_H100_v5	H100 SXM	8	96	1,900 GB	8x 3.84 TB	3,200 Gbps IB
ND96amsr_A100_v4	A100 80GB	8	96	1,900 GB	8x 1.9 TB	1,600 Gbps IB
a3-megagpu-8g	H100 SXM	8	208	1,872 GB	6 TB SSD	GPUDirect-TCPXO
a2-ultragpu-1g	A100 80GB	1	12	170 GB	375 GB SSD	Up to 100 Gbps

Neocloud and marketplace providers that price per-GPU-hour directly do not involve this simplification — their assessed price is the quoted GPU price with no bundling adjustment. The bundle composition table above applies only to hyperscaler assessments where instance-level pricing is the input.

5. Benchmark Calculation

5.1 Headline Statistic: Median

The published headline value for each benchmark series is the median (P50) of the assessed prices from all contributing providers within that source family.

The median is chosen over the mean because it is robust to outliers and avoids giving disproportionate weight to a single provider's pricing decision. It is also easier to defend as a neutral

reference point than a trimmed mean or percentile.

5.2 Supporting Statistics

In addition to the headline median, Pulse calculates and stores the following for each benchmark series on each assessment date:

- **P25 (25th percentile):** A competitive-price indicator. Stored for internal use and may be published separately as a "Pulse Competitive Price Indicator" in future versions.
- **P75 (75th percentile):** The upper end of the assessed range.
- **Min / Max:** The lowest and highest assessed prices.
- **Provider count (n):** The number of providers contributing to the assessment.
- **Individual provider assessments:** Each provider's assessed price is stored individually with full provenance (raw collection ID, collector version, methodology version, normalization notes).

5.3 Weighting

Within a benchmark series, all contributing providers receive equal weight. No volume-weighting, market-share-weighting, or tiering is applied. This is the simplest approach to defend and explain, and avoids introducing subjective capacity estimates into the calculation.

Equal weighting means the benchmark reflects the median available list price, not the volume-weighted transaction price. This is a deliberate design choice: Pulse measures posted pricing (analogous to rack rates), not cleared volume.

5.4 Calculation Example

For the Pulse H100 SXM Hyperscaler On-Demand assessment on a given date:

1. Collect H100 SXM on-demand prices from AWS, Azure, and GCP
2. Normalize each to USD/GPU-hour using provider-specific rules (Section 4.4)
3. Sort the four prices: e.g., $6.88 (AWS), $10.00 (OCI), $10.98 (GCP), $12.29 (Azure)
4. Median = average of two middle values = $10.49
5. Publish: Pulse H100 SXM Hyperscaler On-Demand = $10.49/GPU-hour

6. Publishability Criteria

Not every benchmark series is published on every day. To ensure that published assessments are meaningful and defensible, each series must meet source-family-specific publishability thresholds. These thresholds differ by family because each family represents liquidity differently.

6.1 Hyperscaler Thresholds

- **Publishable:** 3 of 4 major hyperscalers report a valid price
- **Ideal:** 4 of 4 hyperscalers report a valid price
- **Unpublishable:** Fewer than 3 hyperscalers report

Rationale: The hyperscaler universe is exhaustive at n=4 (AWS, Azure, GCP, OCI). Three of four hyperscalers represents sufficient market coverage for a meaningful assessment. With n=4, the median is the average of the two middle provider prices, providing a more robust central tendency than the n=3 median which structurally pinned to a single provider's price. The full provider-level price set (P25, median, P75, and individual contributing prices) is published alongside each assessment to ensure transparency.

6.2 Neocloud Thresholds

- **Publishable:** 3 or more neocloud providers report a valid price
- **Caveated:** 2 providers report (published with "thin data" caveat)
- **Unpublishable:** Fewer than 2 providers report

Rationale: The neocloud universe is larger and more dynamic. Requiring at least 3 providers ensures the median is not determined by a single provider's pricing decision. With 5 active neocloud sources for both H100 SXM and A100 80GB, these thresholds are comfortably met.

6.3 Current Coverage Status

Benchmark Series	Providers Contributing	Count	Status
H100 SXM Hyperscaler On-Demand	AWS, Azure, GCP, OCI	4	Publishable
H100 SXM Neocloud On-Demand	Lambda, RunPod, CoreWeave, Paperspace, DataCrunch	5	Publishable
A100 80GB Hyperscaler On-Demand	AWS, Azure, GCP, OCI	4	Publishable
A100 80GB Neocloud On-Demand	Lambda, RunPod, CoreWeave, DataCrunch	4	Publishable

7. Data Quality and Anomaly Handling

7.1 Automated Validation

Each collected data point undergoes validation before inclusion in a benchmark assessment:

- **Non-null price:** Prices of zero or null are excluded.
- **GPU model resolution:** Only prices that map to a canonical GPU model are included. Unrecognized GPU names are logged and excluded.
- **Region filter:** Only US-region pricing is included for hyperscalers. Neocloud providers that publish a single global price are included without region filtering.

• **Pricing type classification:** Only on-demand and spot prices are included. Committed-use, reserved, calendar-mode, and defined-duration pricing are excluded from all benchmark series.

• **Hyperscaler SKU deduplication:** Hyperscaler pricing APIs may return multiple SKU variants for the same instance type, differing by operating system, tenancy model, capacity reservation status, or bundled software. Each collector must filter to exactly one canonical SKU per instance type that represents the standard on-demand Linux price. The specific filter attributes vary by provider (see Section 4.2.1) but the principle is uniform: one instance type produces one assessed price.

• **VRAM-based disambiguation:** For marketplace providers where multiple VRAM variants share the same GPU name (e.g., Vast.ai listing both 40GB and 80GB A100s as "A100 SXM4" or "A100 PCIE"), the collector must check the VRAM field to assign the correct canonical GPU model. Listings with VRAM ≤ 45,000 MB are classified as the 40GB variant.

7.2 Anomaly Review

When a provider's assessed price deviates by more than 50% from the current benchmark median for that series, it is flagged for review. Flagged prices are included in the benchmark unless determined to be a data classification or source-quality issue (e.g., API error, stale data, misclassified SKU). The default is inclusion; exclusion requires a documented rationale.

Any exclusion decision is recorded in the Data Changelog with the affected date, provider, and reason.

7.3 Missing Data

If a provider's API is unreachable or returns an error on a given collection day, that provider's data is marked as missing for that date. The raw assessment database is never forward-filled or interpolated—missing data stays missing in the underlying data layer.

However, for the purpose of benchmark series calculation (index formation), a staleness carry-forward policy applies. When a provider did not contribute a fresh assessment on the current assessment date, the benchmark query will use that provider's most recent valid assessment from the preceding 3 calendar days. This prevents a temporary API outage from artificially reducing the provider count and distorting the benchmark median. The carry-forward rules are:

● **Staleness window: 3 calendar days.** If a provider's most recent assessment is more than 3 days old, it is excluded from the benchmark entirely. Three days is sufficient to cover transient API outages and weekend maintenance windows without incorporating stale pricing that may no longer reflect the provider's current rate.

● **Database integrity preserved.** No synthetic or carried-forward rows are inserted into the gpu_price_assessments table. The carry-forward is applied only at the benchmark calculation layer (the SQL query that computes the benchmark series). The underlying data remains a truthful record of what was collected on each date.

● **Transparency.** Carried-forward prices are flagged in the benchmark output. The benchmark query reports each contributing provider's assessed_at date and a boolean

is_carried_forward flag, so consumers can see exactly which providers contributed fresh data and which used a recent prior assessment.

● **Publishability thresholds still apply.** Provider counts used for publishability evaluation (Section 6) include carried-forward providers. If a prolonged outage causes multiple providers to exceed the 3-day window simultaneously, the series may become CAVEATED or UNPUBLISHABLE as normal.

Rationale: Cloud GPU pricing from hyperscalers and neoclouds changes infrequently—typically weeks or months between rate changes. A 1–3 day carry-forward introduces negligible pricing error while preventing an API outage (which conveys no information about the provider's actual price) from reducing benchmark quality. This approach follows standard PRA practice for assessed markets where reporter non-response does not imply a price change.

8. Provenance and Audit Trail

Every assessed price in the Pulse database carries full provenance metadata:

• **raw_collection_id:** Links back to the raw API response stored in the raw_collections table. Raw data is immutable and never modified after collection.

• **run_id:** Identifies the specific collection run that produced the assessment.

• **collector_version:** The version of the collector code that parsed the raw data.

• **methodology_version:** The version of the methodology under which the assessment was made.

• **normalization_note:** A human-readable description of how the raw price was transformed into the assessed price (e.g., instance type, GPU count, component formula).

This provenance chain means any published Pulse assessment can be traced back to the exact API response, parser version, and normalization logic that produced it. If a methodology revision changes the calculation rules, the reprocessing pipeline can rebuild historical assessments from stored raw data under the new rules.

9. Restatement Policy

Published assessments may be restated in the following circumstances:

• **Tier 1 — Correction:** A data quality error is discovered (e.g., a provider's API returned stale data, a normalization bug produced an incorrect price). The affected assessment is recalculated from raw data and the correction is logged in the Data Changelog.

• **Tier 2 — Methodology Revision:** A planned change to the methodology (e.g., adding a new provider to a source family, changing the normalization formula). Historical data may be reprocessed under the new methodology. The revision is documented in the Methodology Changelog and communicated to subscribers in advance.

• **Tier 3 — Structural Change:** A fundamental change to the benchmark structure (e.g., redefining source families, changing the headline statistic from median to mean). Treated as a

new series with a version break.

All restatements are recorded in the Data Changelog (methodology/CHANGELOG.md) with: the affected dates, the original and restated values, the reason for restatement, and the tier classification.

10. Governance and Version Control

This methodology document is version-controlled. The current version is 1.0, effective April 2026.

Changes to the methodology follow a structured process:

- **Minor changes** (adding a new provider within an existing source family, fixing a normalization edge case) increment the minor version (e.g., 1.0 → 1.1) and are logged in the Methodology Changelog.
- **Major changes** (redefining source families, changing the headline statistic, restructuring benchmark series) increment the major version (e.g., 1.*x* → 2.0) and require advance notice to subscribers.

The methodology version is stamped on every assessed price in the database, ensuring that any historical assessment can be traced to the exact rules under which it was produced.

Appendix A: GPU Model Registry

The following canonical GPU models are currently tracked by Pulse. Models marked with an asterisk (*) are included in Version 1.0 benchmark series.

Canonical ID	Display Name	Memory	Architecture	In v1.0 Series
h100_sxm *	NVIDIA H100 SXM 80GB	80 GB	Hopper	Yes
a100_80gb *	NVIDIA A100 80GB	80 GB	Ampere	Yes
h200_141gb	NVIDIA H200 141GB	141 GB	Hopper	No (future)
b200	NVIDIA B200	192 GB	Blackwell	No (future)
a100_40gb	NVIDIA A100 40GB	40 GB	Ampere	No
l40s_48gb	NVIDIA L40S 48GB	48 GB	Ada Lovelace	No
l4_24gb	NVIDIA L4 24GB	24 GB	Ada Lovelace	No
a10_24gb	NVIDIA A10 24GB	24 GB	Ampere	No
rtx_4090	NVIDIA RTX 4090	24 GB	Ada Lovelace	No

Appendix B: Provider Classification

Provider	Source Family	Pricing Model
AWS	Hyperscaler	Per-instance-hour
Azure	Hyperscaler	Per-VM-hour
GCP	Hyperscaler	Component (GPU + CPU/RAM)
Lambda Labs	Neocloud	Per-GPU-hour
RunPod	Neocloud	Per-GPU-hour
CoreWeave	Neocloud	Per-GPU-hour
Paperspace	Neocloud	Per-GPU-hour
DataCrunch	Neocloud	Per-instance (normalized)
Vast.ai	Marketplace	Marketplace listings

Appendix C: Excluded Pricing Types

The following pricing types are collected by Pulse infrastructure but excluded from all Version 1.0 benchmark series:

- **Spot / Preemptible:** Prices subject to supply-based fluctuation and interruption. Coverage is thinner (3–4 providers per GPU model). Spot benchmark series will be considered for Version 1.1 as coverage improves.
- **Committed Use / Reserved Instances:** 1-year and 3-year commitment pricing from hyperscalers. Excluded because these reflect different market dynamics (forward commitment vs. spot availability) and are not comparable to on-demand list pricing.
- **Calendar Mode / Defined Duration:** GCP-specific pricing variants that represent neither standard on-demand nor standard spot.
- **Multi-GPU Commitment Pricing:** Discounted rates for multi-GPU configurations offered by some neoclouds (e.g., Paperspace 8x GPU pricing). Excluded to maintain comparability with single-GPU on-demand rates.

— End of Methodology Document —