Bank of America and Citi Invest in Silicon Data Read the announcement →

H100 Rental Price Over Time (2023–2025): A Complete Market Analysis

Track the dramatic rise and fall of NVIDIA H100 rental prices from 2023 to 2025.



Written by Carmen Li
Founder at Silicon Data

Industry Dec 21, 2025 6 Mins Read

Summarize with AI

H100 Rental Price Over Time (2023–2025): A Complete Market Analysis

Executive Summary

Since its launch in mid 2022, the NVIDIA H100 GPU has undergone one of the fastest and most dramatic price corrections in modern infrastructure history.

Early rental prices routinely exceeded **$7–$10 per GPU-hour**, reflecting extreme scarcity and unprecedented demand from AI labs and hyperscalers. By late 2025, however, the same H100 GPUs across non-hyperscale and marketplace providers are widely available for **$2–$4 per hour**, with spot and secondary markets occasionally dipping even lower. Today's standardized rate is published every day on the H100 Rental Price Index.

At **Silicon Data**, we track H100 pricing continuously across regions and provider types. Using our internal price history—anchored by the uploaded H100 price trend chart—and corroborated by common market pricing from hyperscalers and specialist GPU platforms, this article provides the most complete explanation of **how H100 rental prices evolved over time**, why they dropped so sharply, and what buyers should expect going into 2026.

1. H100 Rental Price Timeline: 2023–2025

The following timeline summarizes the most important pricing milestones, combining Silicon Data's internal tracking with commonly observed public-market pricing.

H100 Rental Price Over Time (Per GPU-Hour)

Period	Hyperscaler	Marketplace	Neocloud	Data-driven context
2023-08 to 2023-12	**$7.62–$7.77**(med **$7.76**)			Only **Hyperscaler** is present, prices are tightly clustered around ~$7.7

Period	Hyperscaler	Marketplace	Neocloud	Data-driven context
2024-01 to 2024-05	**$7.37–$8.24**(med **$7.92**)			Still **Hyperscaler-only**; the monthly average range widens and reaches the low $8s.
2024-06 to 2024-12	**$8.98–$9.39**(med **$9.34**)	**$2.50–$3.19**(med **$2.58**)	**$2.74–$3.01** (med **$2.99**)	**Marketplace starts (2024-06)** and **Neocloud starts (2024-07)**. A clear tiering appears: Hyperscaler ~$9+, Marketplace/Neocloud ~$2.5–$3.0.
2025-01 to 2025-05	**$8.73–$9.30**(med **$8.96**)	**$2.27–$2.46**(med **$2.29**)	**$3.00–$4.19** (med **$3.50**)	Marketplace continues lower (~$2.3). Neocloud shows a wider spread (up to ~$4.19). Hyperscaler remains near ~$9.
025-06 (single month)	$6.94	$2.00	$3.29	A visible **step-down** month for Hyperscaler (from ~$9 to ~$6.94). Marketplace and Neocloud remain near ~$2 and ~$3.3.
2025-07 to 2025-12	**$6.20–$6.64**(med **$6.26**)	**$1.92–$2.00**(med **$1.95**)	**$3.01–$3.40** (med **$3.33**)	Post-shift stabilization: Marketplace clusters near ~$2, Neocloud near ~$3.3, Hyperscaler near ~$6.3. The tiering persists but at lower levels vs. 2024 H2.

2. Why H100 Prices Were So High at Launch

The H100 entered the market under extraordinary conditions. Unlike previous GPU launches, it coincided with:

- Explosive growth in large language models
- A global race to build proprietary foundation models
- Limited early production volumes
- Hyperscalers reserving capacity years in advance

For much of 2023, the effective market for H100 rentals was controlled by a handful of large providers. On-demand pricing from major clouds often exceeded **$12.00 per GPU-hour**,

At Silicon Data, early H100 capacity was allocated almost exclusively to long-term customers, reflecting how constrained the market truly was. Pricing during this phase was not driven by margins alone—it was driven by simple scarcity economics.

3. The Turning Point: Supply, Resellers, and Competition

The price collapse did not happen gradually—it happened structurally.

Increased Supply

By early 2024, NVIDIA significantly ramped H100 shipments. Large colocation facilities in North America and Europe brought entire H100 clusters online within months, dramatically expanding available capacity.

Rise of Compute Resellers

At the same time, a new class of GPU marketplaces and specialist providers emerged. These platforms monetized underutilized reserved capacity, introducing liquidity into a previously rigid market. This created price discovery almost overnight.

Hyperscaler Response

Once specialist providers proved H100s could be profitably rented below $3/hr, hyperscalers were forced to respond. In June 2025, AWS reduced H100 pricing by roughly 30%, triggering a broader market reset.

The result was a flip from **shortage-driven pricing to oversupply-driven pricing**—a classic commodity transition.

4. Regional H100 Price Variations (and Why They Exist)

Even in late 2025, H100 rental prices are not uniform globally.

Typical Late-2025 Pricing by Region across Hyperscale, Neocloud and marketplace

Region	Avg. H100 Price	Primary Cost Drivers
North America	$4.00–$4.50/hr	Scale, power efficiency
Europe	$4.00–$6.20/hr	Energy costs, regulation
Asia Pacific	$3.60–$7.00/hr	Import logistics, demand density
Latin America	$5.50–$10.00/hr	New capacity, lower power costs

These differences are driven by:

- Electricity pricing
- Data center density
- Network transit costs
- Regional demand imbalances
- Policies

For teams with flexible workloads, regional deployment alone can reduce GPU spend by **20–30%**.

5. Spot vs. Reserved vs. On-Demand Pricing Explained

One reason H100 prices appear inconsistent is the existence of multiple pricing models.

On-Demand

- Highest predictability
- No interruptions
- Typically 20–40% more expensive

Reserved / Committed

- Long-term commitment (30 days to 1 year)
- Lower effective hourly rate
- Best for sustained training workloads

Spot / Preemptible

- Up to 60% cheaper at times
- Risk of interruption
- Ideal for fault-tolerant jobs, fine-tuning, batch inference

6. Why Prices Dropped So Fast: Deeper Market Mechanics

Beyond supply increases, several less obvious forces accelerated the price decline:

Oversupply from Overcommitment

Many providers overestimated long-term training demand in 2023. As fewer new foundation models launched in 2024–2025, excess capacity flooded the market.

Shift from Training to Inference

While model training drove demand in earlier years, 2025 saw a clear shift toward **inference-centric workloads** as organizations focused on deploying models at scale.

However, inference isn't monolithic:

- **Real-time inference** (e.g., chatbots, search, retrieval) requires consistent availability and low latency, favoring **reserved or on-demand compute** to ensure SLAs.

- **Batch and asynchronous inference** (e.g., embedding pipelines, bulk summarization) is **interruptible and queue-based**, making it well-suited to **spot or auction-priced GPUs**.

Improved GPU Efficiency

Software optimizations reduced GPU-hours required per model, effectively lowering demand without reducing compute output.

Together, these forces created a buyer's market in record time.

7. Risks and Considerations for H100 Renters

Lower prices do not eliminate risk. Buyers should still watch for:

- **Spot interruptions** on volatile markets
- **Hidden fees** (egress, storage, networking)
- **Hardware variance** (80GB vs. lower-memory SKUs)
- **Regional quotas or waitlists**
- **Opaque pricing that changes post-deployment**

At Silicon Data, we mitigate these risks through real-time pricing visibility, hardware standardization, and historical cost tracking.

8. Renting vs. Buying an H100: The Economics

A common question is whether it makes sense to buy H100 hardware outright.

At **$3/hr**, running an H100 continuously costs:

- ~$72/day
- ~$2,160/month
- ~$26,000/year

That roughly equals the **purchase price of the GPU alone**, without factoring:

- Power
- Cooling
- Rack space
- Maintenance
- Downtime risk

For most teams, renting remains the more flexible and capital-efficient option—especially as prices continue falling.

9. Silicon Data's Position in the H100 Market

Silicon Data was built specifically to address the volatility exposed by the H100 cycle. Our platform provides:

- Real-time and historical H100 pricing
- Region-aware cost optimization
- What we internally call the **Silicon H100 Rental Index**—a normalized benchmark tracking price movement over time

This allows customers not just to rent GPUs, but to understand **when, where, and why** to do so.

Conclusion: From Scarcity to Strategy

The story of**H100 rental price over time**is ultimately a story of market maturity. What began as an exclusive, high-cost resource in 2023 has become a competitive, transparent commodity by late 2025.

For buyers, this shift creates opportunity—for those who understand the forces behind the numbers. With proper pricing models, regional awareness, and provider transparency, H100 compute is now more accessible than ever.

At Silicon Data, we believe the next phase isn't about cheaper GPUs alone—it's about **smarter GPU economics**.



Written by Carmen Li
Founder at Silicon Data

Share this story

Subscribe to our Newsletter

Your email

Subscribe

Articles you may like

Steve Hou • Nov 16, 2026

Hopper and Ampere rents are rising, but the 30–50% is a reset, not a quarter

CoreWeave said Ampere and Hopper pricing rose 30–50% quarter over quarter. Our indices rose 2–13% in Q3, but measured from 2025 the H100 index sits inside that range.

Steve Hou • Nov 11, 2026

The Crunch Spread: GPU Rents and Token Prices Are Pulling Apart (Part 1)

GPU rents are rising and token prices are falling. We measure the crunch spread as break-even throughput: the tokens per second a GPU must serve to cover its rent.

Steve Hou • Oct 21, 2026

Dynamic Pricing Brings Price Discovery to GPU Spot Rental Markets

More than half of GPU spot price series did not change once in nine days. Nebius's new dynamic

Silicon Data • Sep 24, 2026

B200 Rental Price Update, August 2026: The Quietest Month Yet

B200 rental index at 5.62 entering September, up 27.6% YTD: the quietest month yet, a term curve

spot pricing shows what a price that moves with capacity reveals.

→

below spot, and B200 futures listing in October.

→

Silicon Data • Sep 15, 2026

A Compute Provider's Guide to Arbitrage: When Reserve Contracts and GPU Futures Disagree (Part 2)

How to read a GPU reserve contract against the compute futures strip: the implied forward rate, the size of a gap, and why the two markets differ.

→

Silicon Data • Sep 14, 2026

The Practitioner's Guide to the Compute Future: How to Hedge, Step by Step (Part 1)

How buyers and providers can lock in H100 rental costs with the new NYMEX compute futures: contract facts, a worked 50-GPU hedge, daily cash flows, basis risk.

→

Make better compute decisions today

The data intelligence platform delivering systematic visibility and institutional benchmarks across the AI economy.

Getting Started | Talk to Sales

Subscribe to our newsletter

Research and market updates on GPU compute pricing, straight to your inbox.

Your email | Subscribe

Products

- H100 Index
- H200 Index
- A100 Index
- B200 Index
- B300 Index
- MI300X Index
- SiliconMark™
- SiliconSiteIQ™

Indices

- GPU Rental Price Indices
- H100 Index
- H200 Index
- A100 Index
- B200 Index
- B300 Index
- MI300X Index
- RAM Index

SiliconPriceIQ™

SiliconCarbon™

GPU Pricing Dataset

Token Market Pulse

Token Pricebook

Index & Curve Datasets

GPU Forward Curve

GPU Residual Value

LLM Token Index

Company

About us

Media

Pricing

Contact Us

Resources

Blog

Use Cases

Events

Market Research

Documentation

Terms of Use | California Privacy Notice | Privacy Notice | Report a Concern | Resources

© 2026 Silicon Data® is a registered trademark of Silicon Data Inc. All rights reserved.

Ask AI for a summary of Silicon Data