The Great GPU Shortage – Rental Capacity – Launching our H100 1 Year Rental Price Index

H100 Rental Prices up 40%, GPU Rental Pricing Dashboard Launch, Compute Rental Market Structure, Will Rental Prices keep going up?

DANIEL NISHBALL, JORDAN NANOS, CHEANG KANG WEN, AND 2 OTHERS

APR 01, 2026

240 3 27 Share

Anthropic's Claude 4.6 Opus and Claude Code have soared in demand. Anthropic's ARR has more than tripled in just a single quarter from $9B at the end of last year to over $30 today. Open models such as GLM and Kimi K2.5 caused open model use cases to soar. Capital raises by firms like Anthropic, OpenAI, and various Neolabs also demand GPUs.



Claude Code is the Inflection Point

DOUG O'LAUGHLIN, JEREMIE ELIAHOU ONTIVEROS, AND 3 OTHERS • FEB 5

Read full story →

This inflection point means that demand has spiked and there's been a run on GPUs at the hyperscalers and Neoclouds.

This new source of demand has spiked pricing for products and services across the supply chain, from DRAM and NAND memory to fiber optic cables, datacenter colocation and gas turbines.

GPU Rental Pricing is the latest of many compute related products and services to see a dramatic tightness in supply and resulting jump in pricing. H100 1-year GPU rental contract pricing has shot up almost 40% from a low of $1.70/hr/GPU in October 2025 to $2.35/hr/GPU by March 2026.

On-Demand GPU rental capacity is sold out across all GPU types – those that have locked up on-demand instances are not willing to relinquish this capacity back into the pool despite recent price hikes. Trying to find GPU compute in early 2026 has been

like trying to book airplane tickets on the last flight out, high prices, and almost no availability. That's the PC analogy, but the more accurate analogy is that trying to rent a cluster is actually like trying to buy drugs.



Source: SemiAnalysis GPU Rental Price Dashboard

At SemiAnalysis, we have been deeply involved in tracking trends and topics including GPU rental pricing within the Neocloud and Hyperscale ecosystem thanks to our work on ClusterMAX, InferenceX and AI Cloud Total Cost of Ownership.

We also spend a good amount of time helping AI Labs connect with Neoclouds and find GPU rentals in the market and have been actively discussing GPU rental price trends with nearly everyone in the ecosystem.

Since 2023, for our clients, we have maintained GPU rental price indices tracking pricing for most major GPU types (H100, H200, B200, B300, GB200, GB300, MI300, MI325, MI355) across all key rental terms, from on-demand and 1 month all the way to 5 years. Our index is constructed using survey data polling many Neoclouds and buyers of compute which is also validated by transaction data as well as by negotiations and transactions we ourselves are facilitating.



Source: SemiAnalysis GPU Rental Price Dashboard

Today, we are making our H100 1 year GPU rental index publicly available to provide additional data and insights to the community. We update the index monthly and will be providing commentary on the latest trends and anecdotes via our social media pages on X and LinkedIn. Access to the full-term structure and rental pricing across all other major GPU types is available to institutional subscribers of our AI Cloud TCO model.

Click here to access our GPU Rental Index Dashboard

This report will discuss the latest trends, anecdotes and data points regarding the GPU rental market, explain how we analyze the broader GPU Rental market structure and touch on what the future may hold for rental pricing.

Surge Pricing Comes to the GPU Rental Market

The chart illustrating the 1y H100 rental price hardly does the trend justice – anecdotes from our first-hand experience in trying to procure compute and feedback from others in the market paints an even starker picture.

Demand is strong across many very heterogeneous use cases and there is no one-size fits all approach. There are plenty of inference workloads like large mixture of experts (MoE) inference that run best on the latest large world-size systems like the GB300 NVL72, while training workloads can have the best price performance on H100s, keeping demand high even for older cards.



InferenceX v2: NVIDIA Blackwell Vs AMD vs Hopper - Formerly InferenceMAX

DYLAN PATEL, CAM QUILICI, AND 5 OTHERS • FEB 16

Read full story →

Customers are fighting to pay $14/hr/GPU for p6-b200 spot instances in AWS, some Neocloud Giants no longer sell single nodes, H100s are getting renewed at the exact same rate they were signed at 2-3 years ago and some H100 contracts are being renewed for 4 years though 2028. Hunting for even 8 nodes (64 GPUs) of H100s or H200s is not easy – half the providers we asked were completely sold out, and most providers will simply respond they have no capacity of Hopper GPUs coming off contract at all.

We have even heard of renters of compute subdividing their clusters and subletting the compute just like an apartment during the Monaco Grand Prix. Coming soon – the rise of Neocloud slumlords?

Blackwell availability is very tight too. We are hearing lead times for new Blackwell deployments now extending into June-July thanks to strong demand for open-weight models as well as the ongoing surge in inference demand, and most of these clusters are now getting taken up. Indeed, market-wide, all capacity coming online until August to September 2026 has already been booked!

GPU Rental Prices – The Comeback Kid

But how did the market come to this point? Only six months ago, most market observers were skeptical on GPU terminal value and assumed an inexorably steep fall in GPU rental rates over time. Financial analysts chastised any Neocloud or Hyperscaler that used a 6-year depreciation period for its GPU compute assets. Let's quickly recap the story so far before we discuss how trends could evolve in the future.

Before late 2025, the prevailing expectation across the ecosystem was that Hopper (i.e. H100 and H200) rental prices would drop considerably as Blackwell deployments ramped given the latter's much lower cost of compute. Instead, the opposite happened in late 2025: demand for H100s was holding firm, and in many cases, *strengthening*. The rapid adoption of open-weight models and accelerating inference demand at that time was the first sign of the insatiable wave of compute demand coming to market.

January was the next inflection point for compute when memory pricing, across both DRAM and NAND pricing, went from rising aggressively for several quarters, to going completely parabolic, with LPDDR5 and DDR5 contract prices tracking toward ~4x and ~5x year-on-year increases respectively in 1Q26 based on our Memory Model.

To manage margin risk stemming from this rapid hike in component costs, OEMs began repricing AI servers at levels that significantly exceeded the underlying increase in component costs. This complicated the cluster capital investment processes as higher server acquisition costs compressed prospective project returns, forcing some operators to slow-roll or abandon deployments. In effect, supply that would have otherwise come online was being withheld, tightening the rental market further.

B300 Liquid Cooled Chassis BoM									
		September BoM		December BoM		Current BoM		4Q26 BoM	
Component + Warranty	Quantity	Unit Cost	Extended Price	Unit Cost	Extended Price	Unit Cost	Extended Price	Unit Cost	Extended Price
CPU DDR5 RAM (per GByte)	3,072	$4.0	$12,288	$8.0	$24,576	$14.0	$43,008	$21.7	$66,601
Local NVMe Storage (per TByte)	30	$80	$2,400	$120	$3,600	$156	$4,680	$156	$4,680
Total Cost									
% Difference vs Sept					*3%*		*8%*		*14%*

Source: SemiAnalysis AI TCO Model

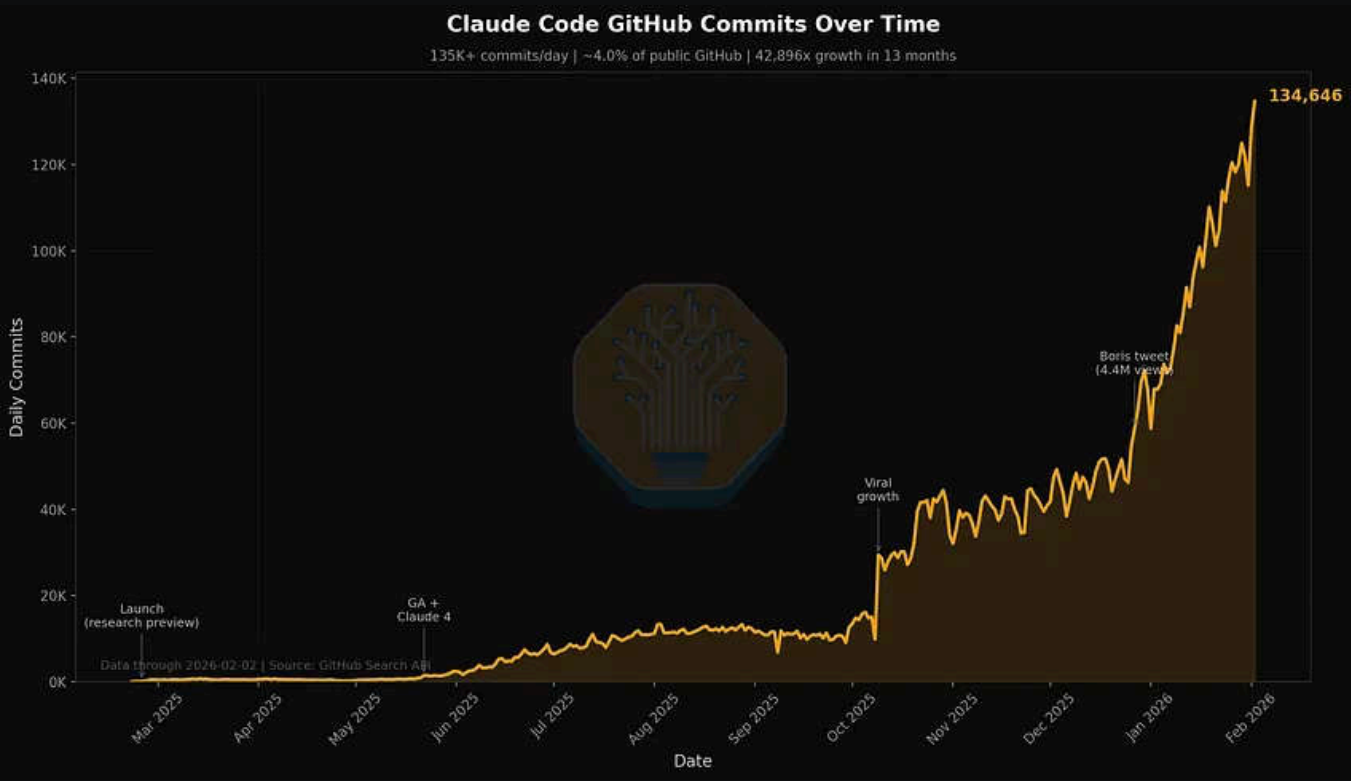
Amid the server procurement disarray cause by this AI Server Pricing Apocalypse, GPU rental demand was clearly accelerating, with most remaining spare capacity taken up completely during January and February. By March – it became increasingly impossible to find any H100s, H200s or B200 rental capacity for any term. Rental pricing broke above $2/hr/GPU for a 1y contract by late January, and then shot up 15-20% by mid-to-late February vs end January and is set to rise another 15-20% month-on-month by the end of March.

A major driver of demand early this year arose from native media generation - Seedance and Nano Banana are driving massive increases in token throughput as users generate and refine images and video at scale. But the most visible driver of demand is the emergence of multi-agent workloads executing multi-step workflows, operating at high concurrency and iterating continuously, leading to parabolic growth in token and compute consumption.

Look no further than the trends regarding Claude Code that we have already called out in many articles. SemiAnalysis as a company has, over the past 7 days, consumed billions of tokens costing around ~$5/M tok on average, but the return on time saved and expansion of workflows and capabilities far exceeds that cost. SemiAnalysis now deploys a suite of AI tools across workflows beyond simple search and summarization – notably dashboarding, automated scraping, large-scale data wrangling and agentic

financial modelling.

We are tracking the overwhelming demand through proxies like Claude Commits Daily. At the current trajectory, we believe that Claude Code will be 20%+ of all daily commits by the end of 2026. While you blinked, AI consumed all of software development. Institutional clients who are interested in the data set can reach out to our API documentation team. A spoiler alert: this number of commits is meaningfully higher than when we published.



Source: SemiAnalysis Tokenomics Dashboard, Contact sales@semianalysis.com for data access

It seems that almost everyone in our circle is an avid user of Claude Code – yet we recognize that our circle of contacts are immersed in everything AI and Semiconductors, and thus they are just the tip of the spear. For many Fortune 500 companies and to the broader world, Claude Code and the Agentic world are but a quirky side story that may come up on their Facebook feeds or their favorite NPR podcasts. They are completely unaware of the tidal wave of productivity and upheaval that the agentic world is about to unleash upon the world.



Source: SemiAnalysis, Generated by Google Gemini

As those in the broader economy that recognize the amazing return on investment from using AI Tools join us in surfing this tidal wave, there will continue to be a step shift higher in token consumption over time. The debate on the true return of using AI is now a settled question – the use of AI tools can deliver value an order of magnitude greater than the cost of using the tools. The shift up and to the right in the demand curve for tokens is providing a powerful and relatively inelastic (for now) force driving up GPU rental pricing.

Put simply – if the return on investment from using AI tools is 5-10x, then there is clearly a long way to go in GPU rental pricing before prices rise enough to curtail demand. It would not surprise us for the increases in rental pricing to exert further upward pressure on server and component costs.

Introducing the SemiAnalysis 1Y H100 contract price index

Today, we are making our SemiAnalysis H100 1-year rental contract price index freely available to create greater awareness and transparency around trends in the GPU rental market.

[Click here to access our GPU Rental Index Dashboard]

Our index is constructed from direct survey data across a pool of 100+ market participants including Neocloud providers, buyers and sellers of compute that is captured every month to determine a representative range (25th to 75th percentile) for GPU rental contracts. We validate these pricing levels with transaction data as well as by arranging a few transactions ourselves, connecting buyers and sellers of compute within our network.



Source: SemiAnalysis GPU Rental Price Dashboard

Since 2023, we have tracked the contract market price across 3m to 5y tenors for the H100, H200, B200, B300, GB200, GB300, and we have selected data available for the AMD complex as well (MI300, MI325, MI355).

The SemiAnalysis H100 1-year contract price index is additive to currently available GPU indices for a few different reasons:

- Many GPU rental indices are derived from spot/on-demand listings or posted pricing, but most of the GPU rental market is transacted on a long-term basis with contracts of at least 6mths and longer. These prices are negotiated on a bilateral basis and such quotes and transactions prices are not typically posted to any freely available online databases. Most large Neoclouds have a preference to rent out capacity on at least a 1-year term with 2 or 3 year terms preferred – even better if they can land 5y large offtake agreements. The SemiAnalysis H100 1-year rental contract price index specifically targets the contract market, where most of the rental volume is transacted. The index references a specific tenor, making it easier for users to understand what market segment it addresses and allowing users to validate these trends against what they may be seeing.
- There is no guarantee that buyers of compute are actually transacting at prices publicly posted by Hyperscalers and Neoclouds. These posted prices may shift around, giving a helpful directional signal for how GPU rental prices are trending but look at these shifts cannot provide an accurate estimate for actual transaction prices. Often times, these publicly posted prices only adjust after the contract market has shifted, lagging actual trends in compute demand. In particular, the on-demand operates by fixing price at a constant level with take-up rates or utilization rates the variable. This pricing is only adjusted on an ad-hoc basis as needed. More on how this market operates later in the article.
- There are many indices that are adept at digesting great volumes of quote, price and transaction data and these can also be great tools for analyzing market trends. By its nature – our approach involves direct interaction with market participants and the people behind these pricing and purchase decisions. There is a story behind every quote and every transaction, and we aim to convey the qualitative and quantitative trends in play as well as anecdotes that are helpful to holistically

understanding the GPU rental market.

Institutional subscribers have access to the full term structure across almost all currently active GPU rental markets.



Current Market Pricing

Latest available pricing across all contract terms ($/hr/GPU)

SKU	Period	On-Demand	1m	3m	6m	1y*	2y*	3y*	5y*
B200	Mar 2026	🔒	🔒	🔒	🔒	🔒	🔒	🔒	🔒
B300	Mar 2026	🔒	🔒	🔒	🔒	🔒	🔒	🔒	🔒
GB200	Aug 2025	🔒	🔒	🔒	🔒	🔒	🔒	🔒	🔒
GB300	Mar 2026	🔒	🔒	🔒	🔒	🔒	🔒	🔒	🔒
H100	Mar 2026	🔒	🔒	🔒	🔒	$2.00–2.70	🔒	🔒	🔒
H200	Mar 2026	🔒	🔒	🔒	🔒	🔒	🔒	🔒	🔒
MI300	Nov 2025	🔒	🔒	🔒	🔒	🔒	🔒	🔒	🔒
MI325	Nov 2025	🔒	🔒	🔒	🔒	🔒	🔒	🔒	🔒
MI355	Nov 2025	🔒	🔒	🔒	🔒	🔒	🔒	🔒	🔒

For full data, contact Sales@SemiAnalysis.com

Source: SemiAnalysis GPU Rental Price Dashboard

Alongside the public release of our H100 1Y contract price tracker, we have launched the SemiAnalysis Tokenomics Dashboard for our institutional Tokenomics model subscribers – a tool for tracking and understanding the frontier AI model landscape. The dashboard lets users build custom benchmarking comparisons across coding, reasoning, math and agentic evals, compare API pricing across leading models and providers, and view key disclosures from major AI labs including token volumes, revenue, valuations and customer counts.

GPU Rental Market Structure Today

Before late 2025, GPU rental pricing was more competitive as operators had much greater GPU inventory while end demand was only starting to accelerate meaningfully. Pricing exercises were competitive, with multiple Neoclouds able to offer very competitive pricing. These operators prioritized utilization and ensuring they did not miss out on sweating fixed compute assets before the next GPU refresh cycle potentially puts pressure on pricing for incumbent GPU servers.

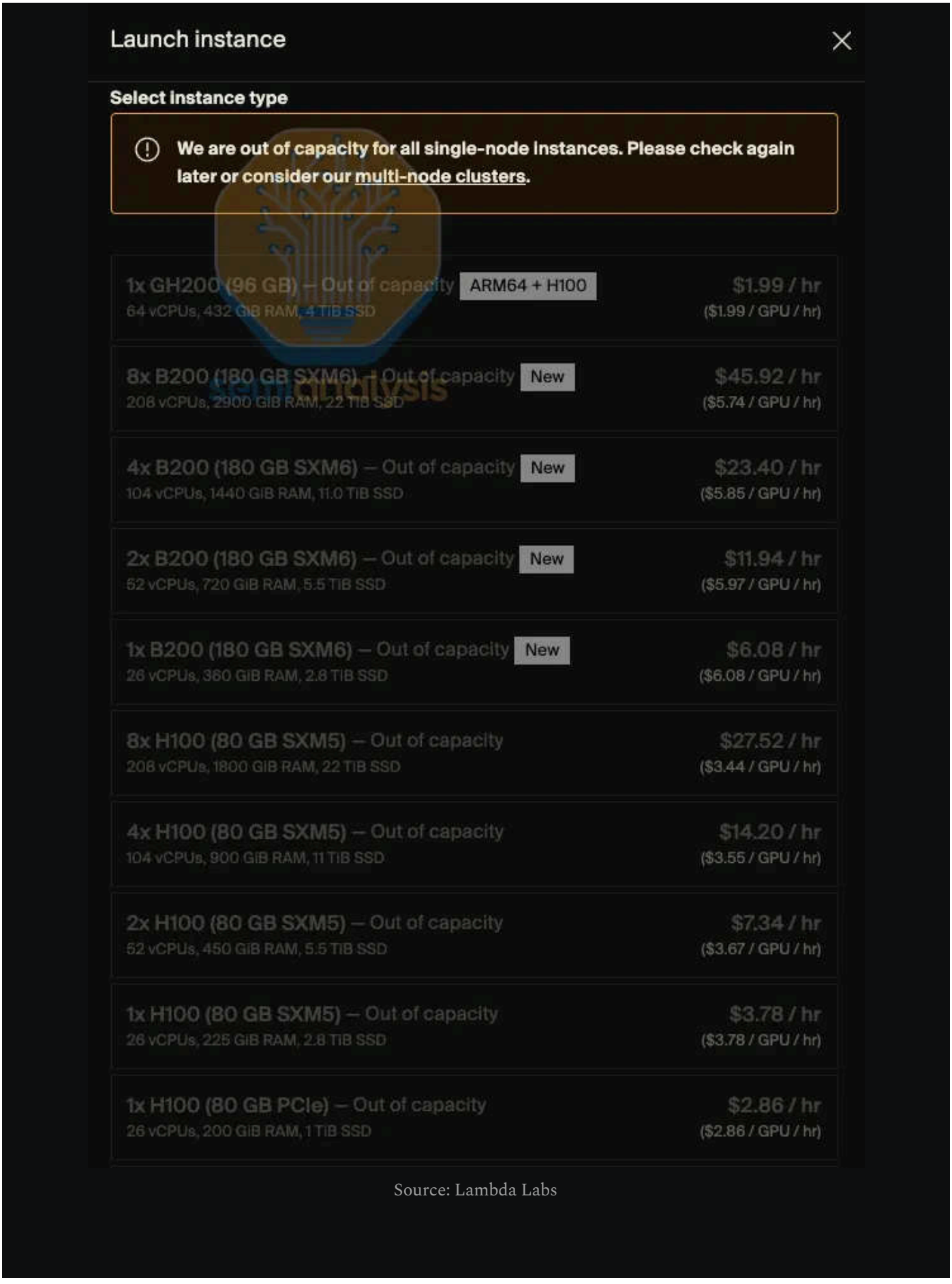
GPU rental providers' strategy has pivoted 180 degrees since then. Neoclouds and Hyperscalers are now in the driver's seat – they can now negotiate for more favorable terms such as higher prepay, better pricing, longer contract lengths and can even pick and choose the contract start and end dates to match their inventory availability. Time is also now on the Neoclouds' side – they can plan deployments on their own time, harnessing the increasing price climate to build the best book of clients for a given cluster over time.

The GPU rental market structure is best understood by dividing it into three primary market segments, each of which tends to cater to different types of customers:

1. Short-term rental: On-Demand, Spot and less than 3-month contracts
2. Mid-term contracts: 3-month contracts all the way to 3-year+ contracts.
3. Long-term offtakes: 4-year to 5-year contracts, though 5-year is the most popular tenor.

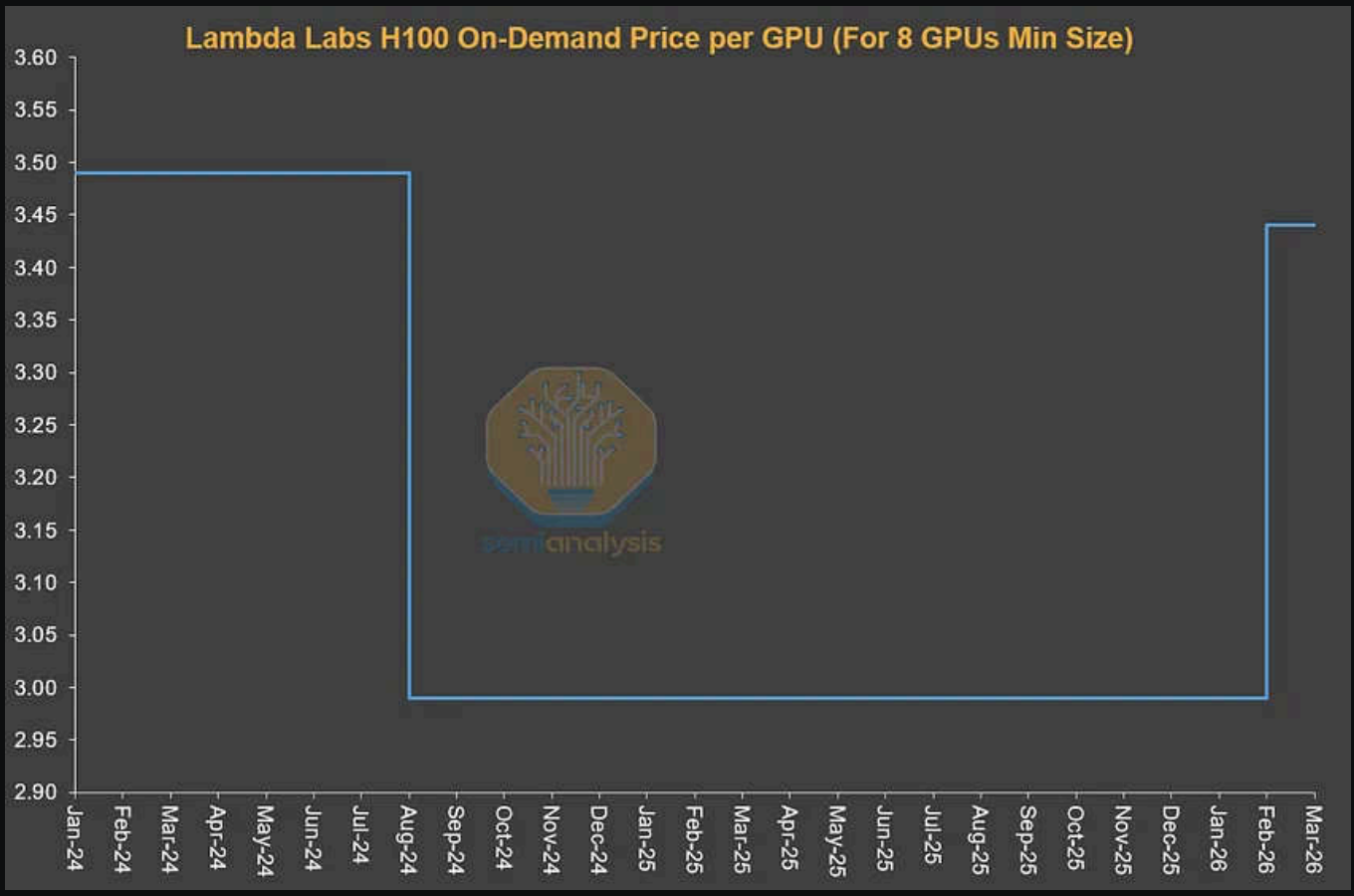
Short-term Rental: On-Demand, Spot, Less than 3-Month Contracts

Short-term rentals represent the very front end of the rental term structure and in many cases represents residual capacity, though many such as Runpod and Lambda very successfully focus on providing considerable capacity of flexible on-demand or spot capacity. On-demand pricing functions tends to function very differently than the rest of the contract GPU rental market. Providers usually set a fixed price for on-demand capacity and will only very infrequently adjust prices.



Source: Lambda Labs

Providers will adjust pricing on a one-off basis in response to utilization levels – if utilization is too low, they will drop pricing to attract demand. If utilization is maxed out – then prices will be hiked as the provider determines that utilization will remain high even at higher pricing levels. This is why a time series of Neocloud posted on-demand pricing will be flat for a long period of time before gapping up or down. With the on-demand market, utilization is the best high frequency indicator of demand, and not price.



Source: Lambda Labs, SemiAnalysis

Mid-Term Contracts

The more economically relevant segments are the contract markets where most GPU rentals by value are transacted. The 1-year segment captures the marginal demand from non-AI lab customers and spillovers from large buyers, making it the most sensitive indicator of tightening conditions. AI Natives and smaller AI Labs can be

seen mostly in the 1-3y tenors, though increasingly these labs and AI Natives are looking to lock up compute by contracting for longer terms of more than 4 years and even agreeing to high prepays of above 20% which were previously atypical for 4y+ deals.

Long-Term Offtakes

Lastly, the 4–5 year segment is dominated by AI labs locking in huge quantities of capacity early. These deals involve large clusters of 50MW or 100MW or even larger – equivalent to about ~24,000 to 48,000 GB300 NVL72 GPUs. In aggregate – these deals represent a very large proportion of the overall Neocloud GPU rental market.

AI Labs like these contracts as they can lock in a huge amounts of compute in one go to cater to rapidly accelerating end demand. The AI Labs also have considerable influence on the cluster design, making decisions regarding storage, networking, CPU compute and more. These are very often bare metal deals as AI Labs have the engineering expertise to customize more layers of the tech stack and extract the highest performance per TCO possible.

Neoclouds like these deals as they can focus their sales resources on just a few large offtake deals rather than dozens of smaller deals which would end up generating the same amount of revenue. Longer-term contracts are also great for Neoclouds as they can then use these contracts to arrange debt financing on favorable terms that will match the tenor of the contract, removing most duration and GPU rental price risk from the equation and locking in a teens project IRR in most cases. It is also common to see Hyperscalers backstop these deals – serving as the direct offtaker from these Neoclouds but then on-selling the compute to an AI Lab. This structure is a win-win for everyone involved – Neoclouds can lend on very favorable terms since their offtaker is a AAA rated Hypercaler, while Hyperscalers can collect a slice of the project revenue by offering the backstop of their balance sheet without actually expanding the balance sheet.

The below table shows a number of large offtake deals we have been tracking. We analyze these deals closely to calculate the implied pricing in $/hr/GPU and profitability metrics like Project IRR, EBIT Margins, etc.

Major AI Cloud Contracts

	Units	IREN-Microsoft	Coreweave-Meta	OCI-OpenAI	Nebius-Microsoft [1]	Coreweave-OpenAI Deal 1	Coreweave-OpenAI Deal 2	Coreweave-OpenAI Deal 3	Nscale-Microsoft Deal 1	Nscale-Microsoft Deal 2[2]	GCP-Anthropic
Announcement Date	Date	3-Nov-25	30-Sep-25	10-Sep-25	8-Sep-25	10-Mar-25	15-May-25	25-Sep-25	16-Sep-25	15-Oct-25	9-Nov-25
Contract Value	USD										
Term Duration (Years)	Years	5	6	5	5	5	4	5	5	5	5
Implied Annual Revenue	USD										
Chip Type	Type	GB300	GB300	GB300 / VR200s	GB300	GB200	GB200	GB300	GB300	GB300	TPU v7
Critical IT Power Contracted[3]	MW										
GPU Capex Disclosed	USD										
All-in Power per GPU	W										
All-in Capex per GPU	USD										
Implied # of GPUs	#										
Deal Revenue per MW per Year	$M/MW/Year										
Annual Deal Revenue per W	$/W										
Implied per Logical GPU Rental Price/Hr	USD										
Prepay %	%										
Depreciation Schedule (Yrs)	Years										
Power Cost	USD/kWh										
Colocation Cost	USD/kW/Mth										
Datacenter Capex	USD M/MW										
Project IRR %	%										
EBITDA Margin	% in Year 1										
EBIT Margin-Depreciation = Deal Term	% in Year 1										
EBIT Margin-Depr. as per Accounting Polic	% in Year 1										
EBIT Margin-6 Year Depr	% in Year 1										

All clusters assume a 3-Layer InfiniBand Network
Prepayment assumptions in red, with prepayments credited equally throughout contract life. Iren prepayment terms is credited to years 3-5. Chip Type assumptions are in red when not disclosed
1. Option to increase contract value to $19.4B
2. Option to add further 700MW in late 2027
3. IREN disclosed 200MW of Crit IT power, with remaining capacity likely reserved for other non-AI compute that supports this cluster in some form

Source: SemiAnalysis AI TCO Model

In today's market climate, most of the large AI Clusters that are ramping are captive to AI Labs, who nonetheless have been dipping into the sub 4y market to supplement existing capacity ramps. They are also preventing supply from re-entering the sub 4y contract market by simply renewing their existing H100 and H200 clusters, indirectly playing in these markets. It will be interesting to see how supply-demand dynamics play out in the 1-3y contract market as AI Labs ramp their GB200 and GB300 megaclusters.

Where The Puck is Going

What is most striking is the disconnect between these underlying dynamics and broader market sentiment. Despite clear evidence of tightening supply and rising prices - conditions that should directly benefit Neocloud providers through margin expansion and stronger arguments for higher useful lives - public market sentiment

has turned increasingly negative on names like CoreWeave, Nebius, IREN, and these companies' share prices are currently at the low end of the 6-12mth trading range.

The market is still anchored to a narrative of eventual oversupply and commoditization, and the developments described above have done little to assuage concerns on GPU terminal value. While, the reality on the ground points to sustained scarcity and pricing power across all Neocloud providers, that suggest that all compute will be in-demand - no matter the relative performance given such an aggressive shortage.

Looking forward, there are three check points to monitor to determine if GPU rental prices will continue to remain elevated.

- GB300 clusters will ramp throughout 2026. We will monitor the extent to which additional compute capacity and thus token volumes coming market ameliorates the ongoing compute crunch or whether token demand will outpace these additions. This will determine the extent to which AI Labs will participate in the sub 4-year market and therefore where pricing trends go for this segment.
- The extent to which the ongoing silicon shortage worsens. We recently wrote about the great AI Silicon shortage, calling out tightness in TSMC's N3 logic wafer capacity and HBM, DRAM and NAND memory among others. This can always get worse as execution hiccups can always arise for any of these complicated manufacturing processes.
- How ARR for AI labs continues to scale – and the rate at which adoption spreads and token consumption continues to grow. Our AI Tokenomics model is squarely focused on this and analyzes key demand and usage signals to track end demand.

Pricing is Only Going One Way For Now, and ROIC Follows

Taken together, these factors point to a clear conclusion: GPU rental pricing is more likely to continue rising than falling.

The dynamic is self-reinforcing. As Neoclouds see supply tighten and prices rise, they move to secure more hardware ahead of further price increases, which only tightens supply and pushes pricing higher still. This echoes the 2023–2024 GPU shortage, when tight supply allowed OEMs to push through outsized margin expansion and drove a sharp spike in server pricing, though we think the server market is mature enough this time around that this may not repeat itself.

As we argued in a recent note to our institutional clients, the re-acceleration in GPU rental pricing improves Neocloud ROIC by expanding margins on already-deployed capital. At the same time, higher rental rates extend the economic useful life of existing GPUs, meaning invested capital generates cash flows for longer before requiring reinvestment.

For now, the clearest beneficiaries are providers with:

- Shorter-duration contracts (repricing faster)
- Large H100 install bases
- Near-term capacity additions

Neoclouds with shorter contract tenors will see capacity roll off and reprice into the current environment, capturing immediate margin expansion. At the same time, Hyperscalers and Neoclouds locking in next-generation capacity over multi-year terms will benefit.

Would we be jinxing it if we said that This Time *Might* Be Different?

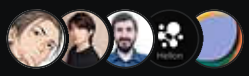
Subscribe to SemiAnalysis

Tens of thousands of paid subscribers

Bridging the gap between the world's most important industry, semiconductors, and business.

Type your email... Subscribe

By subscribing, you agree to Substack's Terms of Use, and
acknowledge its Information Collection Notice and Privacy Policy.



240 Likes · 27 Restacks

← Previous

Next →

Discussion about this post

Comments Restacks



Write a comment...



V3 Apr 2

Long NBIS

LIKE (8) REPLY SHARE



V S Jun 18

You mention: "The dynamic is self-reinforcing. As Neoclouds see supply tighten and prices rise, they move to secure more hardware ahead of further price increases, which only tightens supply and pushes pricing higher still"

But is there another argument - that at some point supply rises more than demand purely for economic reasons (ie customers cant get enough users at specific price points to justify paying for added compute) - what would you say to that out of interest - as net net someone has to pay for it.

LIKE REPLY SHARE

1 more comment...

© 2026 Dylan Patel · Privacy · Terms · Collection notice

Substack is the home for writing, podcasting, and media founders.