AI Value Capture - The Shift To Model Labs

Vera Rubin VR NVL72: V for Value - Rubin delivers a step jump in performance per TCO. ROI accruing to users, Neoclouds, Hyperscalers, AI Labs, Memory Vendors or GPU Manufacturers?

DANIEL NISHBALL, DYLAN PATEL, CHEANG KANG WEN, AND 6 OTHERS

MAY 01, 2026 · PAID

A day in AI now feels like a year in any other industry. Model releases, software breakthroughs, and hardware improvements are compressing multi-year cycles for any other industry into weeks. Over just the past few months, agentic AI has crossed a real inflection point, driving a step-change in the value of tokens while software and hardware improvements have sharply reduced the cost of generating them.

This flood of demand is driven by end users enjoying a huge return on investment (ROI) from consuming tokens, and this demand growth is arguably only in its early innings. This year Anthropic's ARR has exploded from $9B to over $44B today, their gross margins on their inference infrastructure have increased from 38% to over 70% over the same period.

This rapid pace of AI adoption has created value across the stack, but the unique phenomenon is that the AI labs are capturing all the value now, from almost none last year.

End users are enjoying a productivity bonanza, tasks that used to take tens of person-hours costing thousands of dollars can now be accomplished in minutes with a just a few dollars' worth of tokens. This huge surge in revenue and margins is because the value of tokens being created is dramatically improving businesses. For example, SemiAnalysis has reached as high as $10.95 million dollar annual spend rate on Anthropic Claude tokens, but the value we derive allows us to outcompete all our competitors and gain market share.

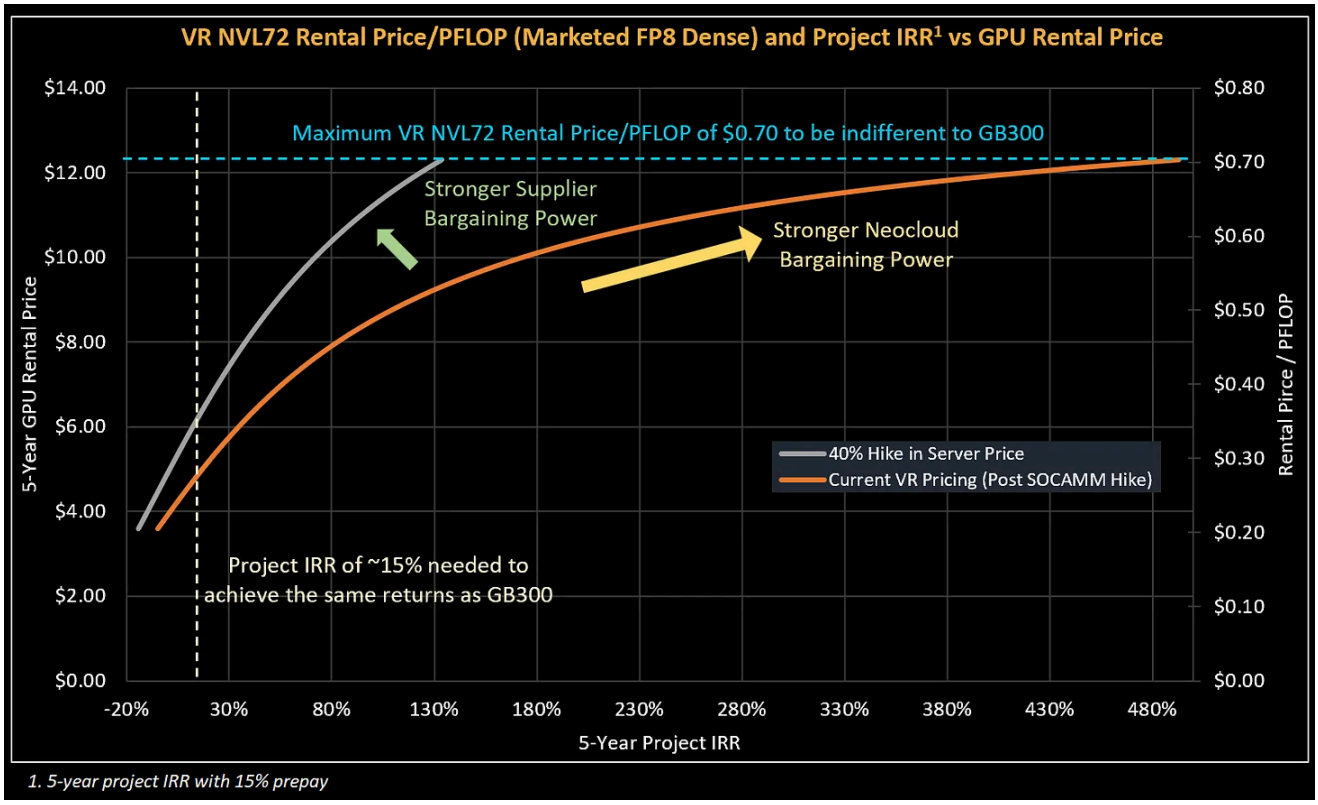
New chips such as Blackwells can generate 30x more tokens per second while running frontier workloads today vs Hoppers a year ago, and ASICs such as TPUv7 and Trainium 3 show similar improvements. Inference providers such as Fireworks, Baseten, Fal, margins are widening while their revenue trends are in hyper growth.

Even parts of the hardware stacks have repriced, with memory prices having gone up 6x in the past year. Neocloud GPU rental pricing is surging as well, up with 1-year H100 rental contract prices up 40% from the bottom in October 2025.

There are two firms in the industry with incredible pricing power that haven't moved much though. TSMC and Nvidia have not reacted to the recent boom in value generation of AI models.

In this article we will explore where value from AI is accruing - from end users to inference providers, Neoclouds as well as hardware providers. We will unveil how TSMC and Nvidia are now venting vast value into every vertical of the ecosystem.

Finally - we introduce a new framework: the "One Chart to Rule Them All" that explores GPU Rental Economics and analyzes whom among the end users, the Neoclouds/Hyperscalers and the AI System suppliers are capturing the most value in the AI ecosystem.



Source: SemiAnalysis AI TCO Model

AI Value Profit Pools

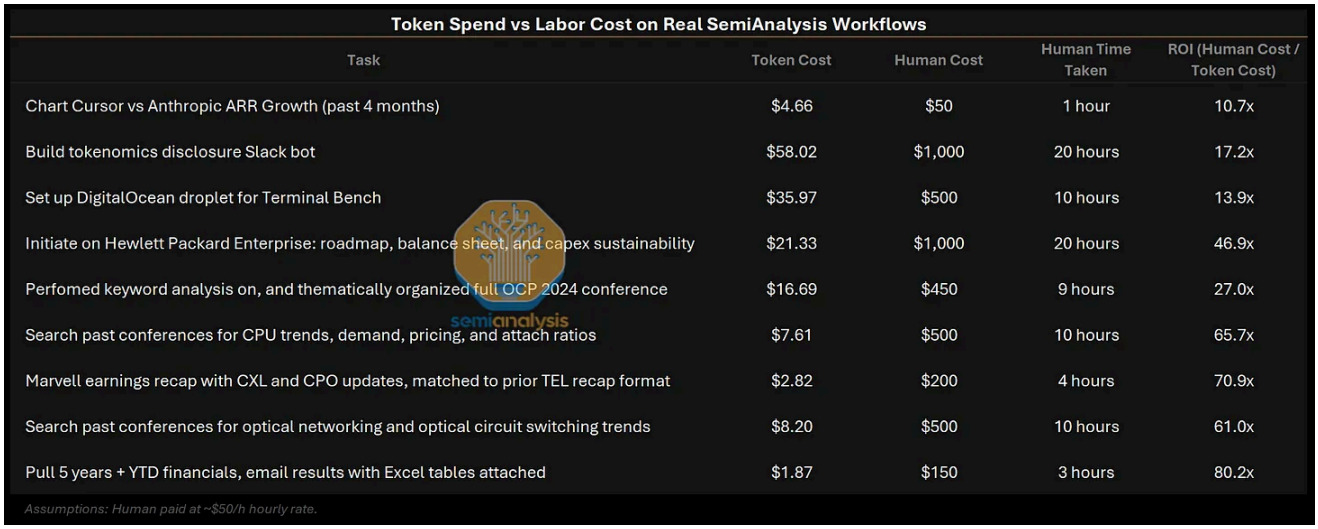
From 2023-2025, all the value in AI was captured by the infrastructure layer. Nvidia had their first blockbuster earnings call in May 2023 and jumped 25% after hours, officially marking the start of the AI trade. In 2024, Vistra and GE Vernova were two of the top performing stocks in the S&P 500 (+265% and +146% respectively) as everyone realized power was becoming the key bottleneck. In 2025, memory stole the show, with SanDisk, Western Digital, Seagate, and Micron all posting 200%+ gains on the year. These are all sweeping generalizations of course and many other infra names have significantly outperformed thanks to increased AI capex. Those interested in all the granular details should subscribe to our institutional products.

During this same period, gross margins for all the model creators and inference providers were famously bad. For most, the actual utility of AI still only amounted to slightly better Google search locked behind a chat interface and Studio Ghibli style selfies. Skeptics loudly proclaimed that there was simply no way AI could ever deliver on the trillions of planned capex.

Agentic AI Has Changed the Game

The world changed in December 2025, when Agentic AI began to *really work*. SemiAnalysis has written and talked extensively about our Claude Code usage, but it is important to emphasize that agentic AI is no longer limited to just coding. Our analysts are using agents every day to convert excel models into dashboards, create charts for all our notes, build financial models and analyze company earnings, and much more. These are all tasks that either 1) we simply wouldn't have been able to do before or 2) would've previously taken our junior analysts many hours, taking them away from far more value added tasks.

The table below shows a handful of real examples from our own workflows, comparing token spend against what the equivalent human labor would have cost:



Token Spend vs Labor Cost on Real SemiAnalysis Workflows

Task	Token Cost	Human Cost	Human Time Taken	ROI (Human Cost / Token Cost)
Chart Cursor vs Anthropic ARR Growth (past 4 months)	$4.66	$50	1 hour	10.7x
Build tokenomics disclosure Slack bot	$58.02	$1,000	20 hours	17.2x
Set up DigitalOcean droplet for Terminal Bench	$35.97	$500	10 hours	13.9x
Initiate on Hewlett Packard Enterprise: roadmap, balance sheet, and capex sustainability	$21.33	$1,000	20 hours	46.9x
Perfomed keyword analysis on, and thematically organized full OCP 2024 conference	$16.69	$450	9 hours	27.0x
Search past conferences for CPU trends, demand, pricing, and attach ratios	$7.61	$500	10 hours	65.7x
Marvell earnings recap with CXL and CPO updates, matched to prior TEL recap format	$2.82	$200	4 hours	70.9x
Search past conferences for optical networking and optical circuit switching trends	$8.20	$500	10 hours	61.0x
Pull 5 years + YTD financials, email results with Excel tables attached	$1.87	$150	3 hours	80.2x

Assumptions: Human paid at ~$50/h hourly rate.

Source: SemiAnalysis

Annualized token spend at SemiAnalysis is already ~30% of employee compensation and we're consuming just under 5B tokens per month per employee (over 5x more than Meta!). This is power law distributed though, so there are team members running over 100B tokens a month. It's obvious that this is still just the beginning, and that all white-collar enterprises will soon embrace agentic AI.

Within the past few months, the value of each token has clearly increased. We estimate that the true blended price per million tokens for running Opus 4.7 on agentic tasks at $0.99 despite the sticker price being $5/$25 per MTok. Agentic workloads have extremely high input-to-output ratios (our Claude Code usage has a ratio of about 300:1) and high cache hit rates (90%+). Because cached input tokens only cost $0.50/MTok, most of the tokens end up in the cheapest tier. We walk through the full methodology here.

When framed this way, it's no wonder why Anthropic ARR has exploded from $9B to potentially $44B+ YTD.

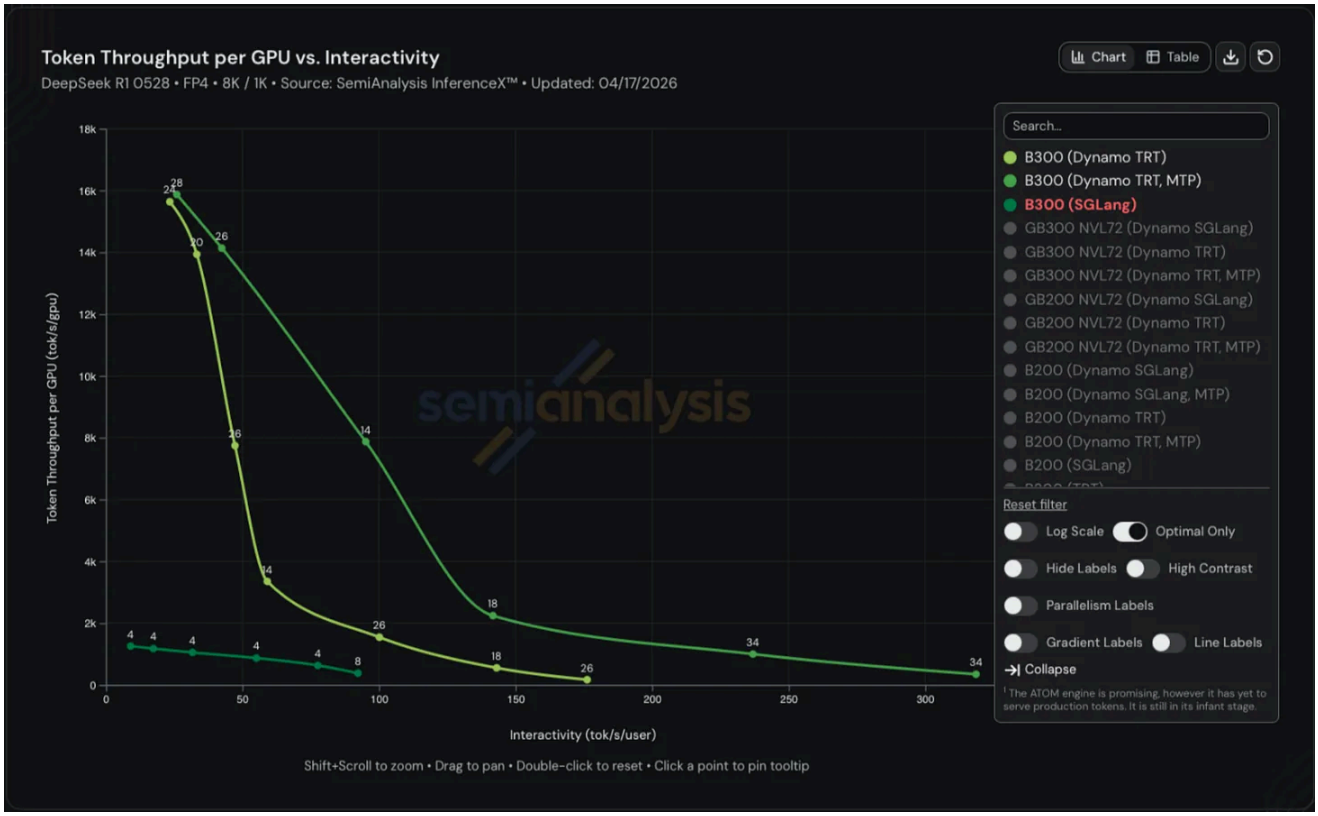
Tokens Are Getting Cheaper to Produce

At the same time, the cost of producing each token has plummeted. This is the largest driver of value accretion to inference providers, and it is a key reason for the sharp increase in margins at large AI Labs.

Cost of production for token has fallen sharply because increases in accelerator pricing generation-over-generation have been more than offset by much higher throughput (tokens/sec/gpu). Average blended price per million tokens has fallen dramatically over the past few months, agentic workloads are inherently multi-turn with longer input/output ratios and higher cache hit rates,but inference margins have gone up from < 40% to > 70% in the same time frame. For in-depth estimates on true blended price per million tokens, token production volumes, and gross margins for all the major models from OpenAI, Anthropic, and more, see our Tokenomics model.

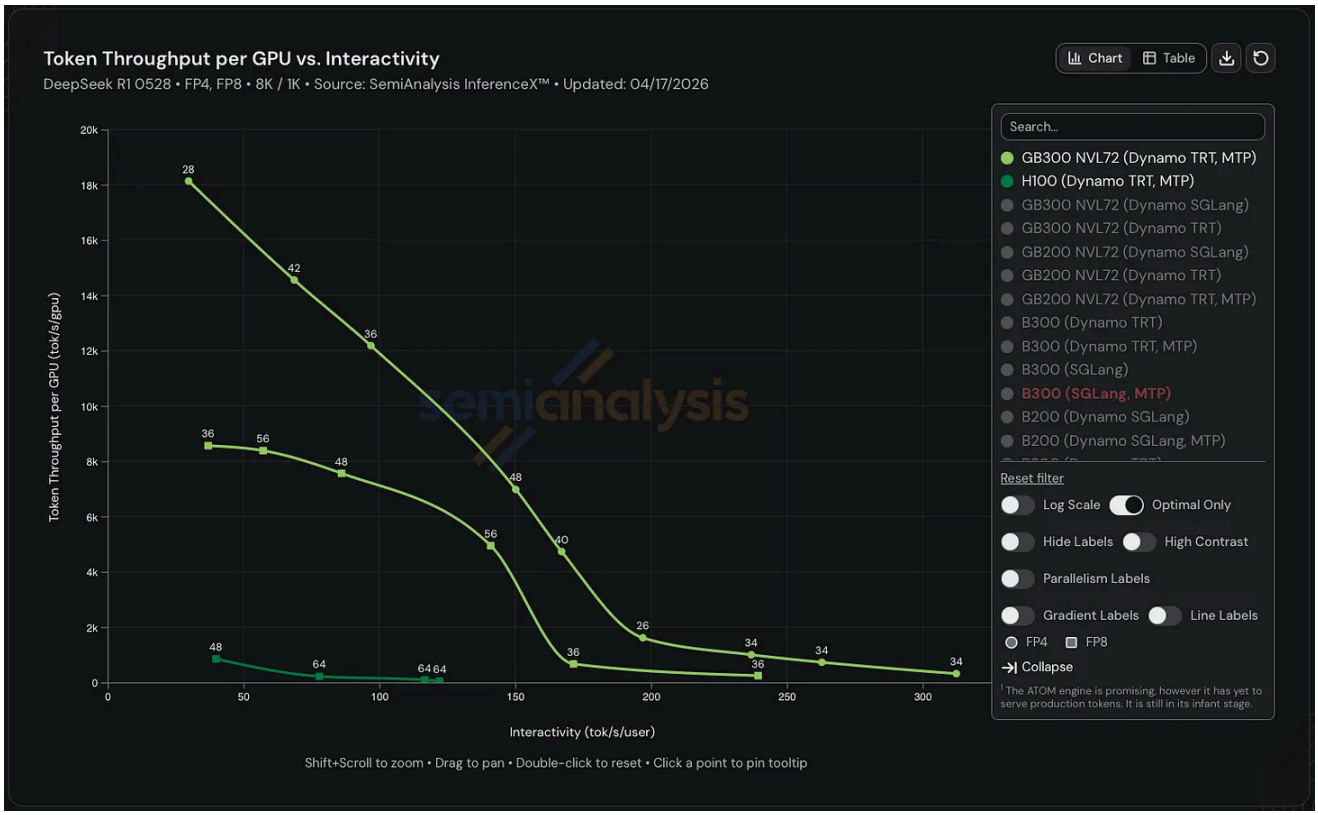
InferenceX remains the best benchmark for tracking real-world inference performance over time for open source models given both hardware and software improvements.

The following chart shows throughput vs interactivity for B300s running DeepSeek R1 on 8k input tokens to generate 1k output tokens. The top line reflects token throughput with wideEP + disagg + MTP, the middle reflects wideEP + disagg and he lowest line is without any of the three software optimizations. The gap is startling with the same B300 able to yield~1k, ~8k, and ~14k tokens/sec/gpu on the same hardware. One can 14x throughput with software improvements alone.



Source: SemiAnalysis InferenceX

If you factor in hardware improvements as well, then the difference is even more pronounced. The most optimized GB300 NVL72 configuration achieves ~17x higher throughput than the most optimized H100 configuration in FP8. If we switch to FP4, which Hopper doesn't natively support, the difference jumps to 32x. Remember that the total cost of ownership per GPU is only ~70% higher for GB300 vs H100.



Source: SemiAnalysis InferenceX

Model Provider Margins Will Continue to Increase

Many were surprised when Anthropic released Opus 4.5 at a price of $5 per million input tokens and $25 per million output tokens in late November 2025. Previous Opus models such as 4 and 4.1 (released May 2025 and August 2025 respectively) were priced 3x higher at $15/$75.

However, we think Anthropic's margins have actually *increased* on Opus tokens despite the lower ASP thanks to software improvements across Trainium and Nvidia GPUs as well as replacing Hoppers with Blackwells.

Anthropic's margin expansion so far has come from cost reductions; they can generate the same tokens for cheaper. Despite the Opus price cut, their ASP/token has also actually gone up because most of the volume shifted from Sonnet to Opus.

Even if XPU providers start dramatically raising prices to better capture their share of the throughput improvements, Anthropic still has another lever to pull to further expand margins: they can continue shifting volume to more expensive SKUs.

As mentioned earlier, the gap between the price of a frontier-level token vs the economic value of the work that can be produced by said token is the largest it's ever been. Anthropic can either re-up the price of the base Opus family or introduce new products. We already saw the latter with Opus fast being priced 6x higher than regular Opus, and Mythos being announced at $25/$125 (5x regular Opus pricing). Both these SKUs are higher margin than regular Opus, yet the most AI-pilled businesses are still more than happy to pay the increased prices because the productivity gains outweigh the cost. If Anthropic let us pay $150/$750 for Mythos fast, we would.

The age of low gross margins for frontier model providers is over. Real agentic AI has permanently increased the market-clearing price per token, and there's no going back.

Why Model Provider Profits Won't Get Competed Away

The most obvious argument for why the labs won't be able to capture higher margins despite increased utility per token is competition. However, we don't think this is how things will play out for two reasons.

First, it's become clear that the frontier model maintains pricing power. Regardless of what the benchmarks may say, open-source models are still noticeably worse than their closed source counterparts for real knowledge work, and there's no reason to believe the gap will close any time soon. Kimi K2.6 ($0.95/$4) exerts very little downward pressure on Opus pricing.

Second, compute constraints means that no single frontier lab will be able to serve the entire market. Anthropic is already beginning to alienate large swathes of the market today by locking Claude Code behind a $100+/month subscription and blocking third party harnesses like OpenClaw. Token demand will far outstrip supply for the

foreseeable future, which means **any lab capable of providing true frontier quality will be able to charge based on the economic value delivered by the token rather than competing away each other's margins.**

Agentic AI Hits the Market, but TSMC and Nvidia Haven't Flinched

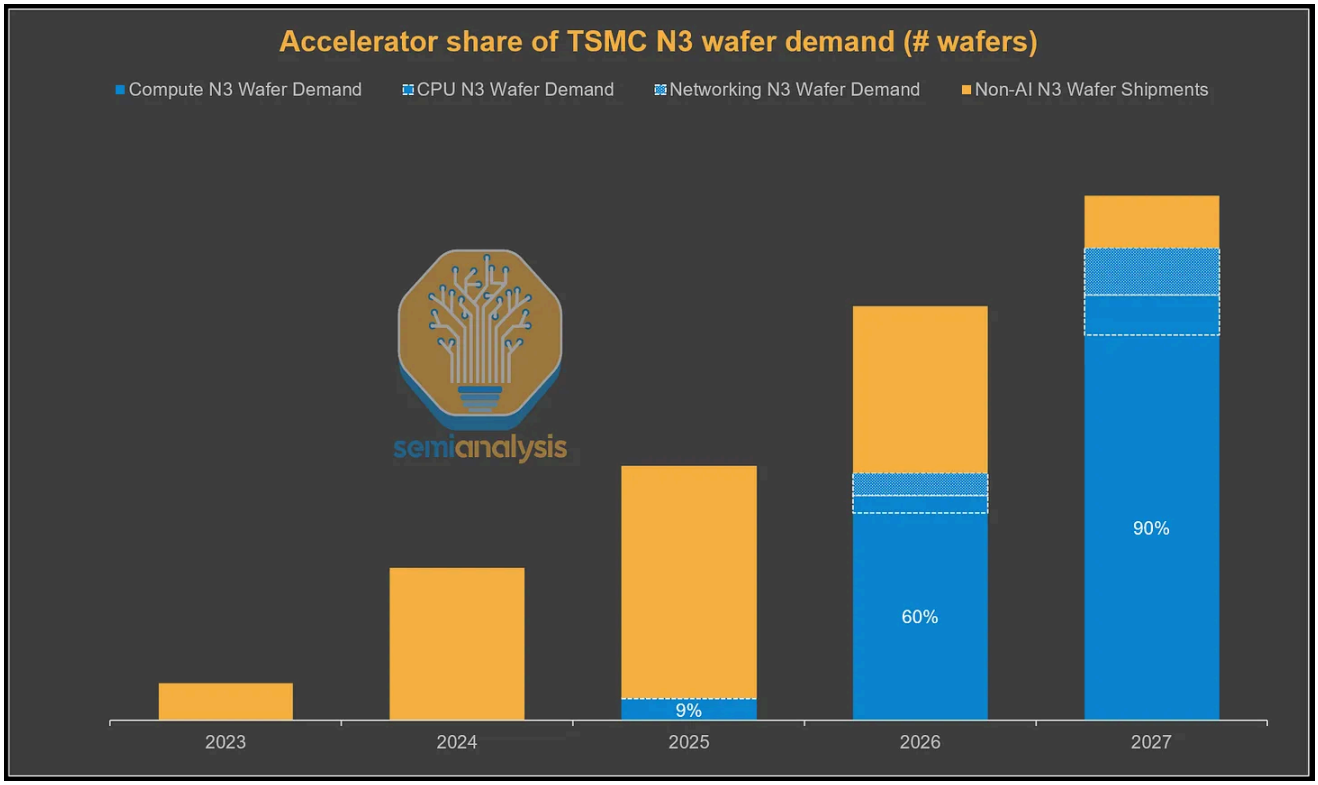
Despite the repeated emphasis on agentic AI during Jensen's most recent GTC keynote, Nvidia and TSMC still have not fully internalized how transformative the past few months have been for token economics. We already saw Nvidia underestimate Blackwell's performance-per-dollar improvements based on Jensen's reaction to InferenceX, and it now appears they have also underestimated how quickly frontier tokens would appreciate in value.

Nvidia is still operating within a framework shaped by prior assumptions, where the willingness to pay per unit of compute declines over time. That assumption no longer holds. The market has shifted materially, driven by the explosion of agentic workloads and a sharp increase in token consumption per workflow. Demand is no longer linear. It is compounding.

Demand, however, continues to accelerate. Anthropic's ARR has reportedly reached $44B+, up from $30B in our last update, while open-weight models such as GLM and Kimi are expanding the addressable compute base. Capital raises across AI labs and neoclouds are translating directly into incremental GPU deployments.

At the same time, compute supply remains structurally constrained. Upstream bottlenecks in memory and leading-edge wafers continue to limit availability, with N3 utilization expected to exceed 100% in the second half of 2026 and DRAM fabs already running above 90% utilization. There is no meaningful relief in sight.

TSMC could raise prices materially, but they haven't. This is a stategic error on their part. If not increasing prices, they could at least demand larger prepayments.



Source: SemiAnalysis Foundry Model, SemiAnalysis Accelerator Model

The current dynamics within the compute market suggest that if current trends continue, the value generated by the overwhelming token end demand will continue accruing to AI Labs, Hyperscalers, Inference Providers, Neoclouds and Memory Vendors.

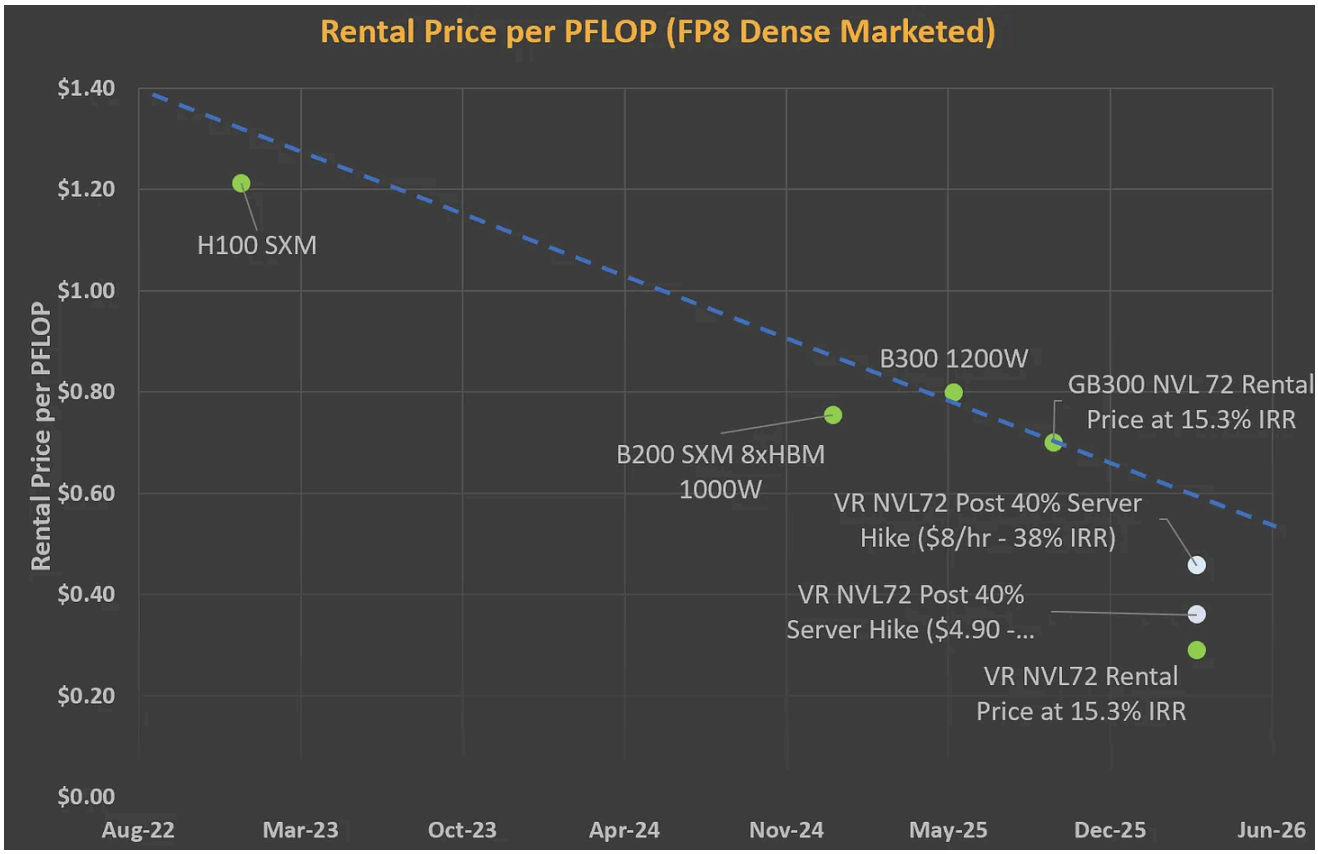
AI labs are capturing a disproportionate share of the value being created, driven by strong end demand, rising token monetization, and increasingly favorable unit economics. At the same time, Nvidia's pricing framework has not fully adjusted to reflect this shift, even as its hardware remains the critical bottleneck enabling that value creation. Despite rising token monetization and increasingly favorable unit economics, Nvidia compute is still the bedrock for enabling that value creation.

Demand for Nvidia systems remains extremely strong across all tiers, with buyers willing to lock in long-term contracts and accept higher pricing to secure capacity. Even with alternative hardware options, Nvidia retains a clear advantage in ecosystem maturity, software stack, and deployment reliability. For many workloads, especially at the frontier, substitutes are not yet fully interchangeable.

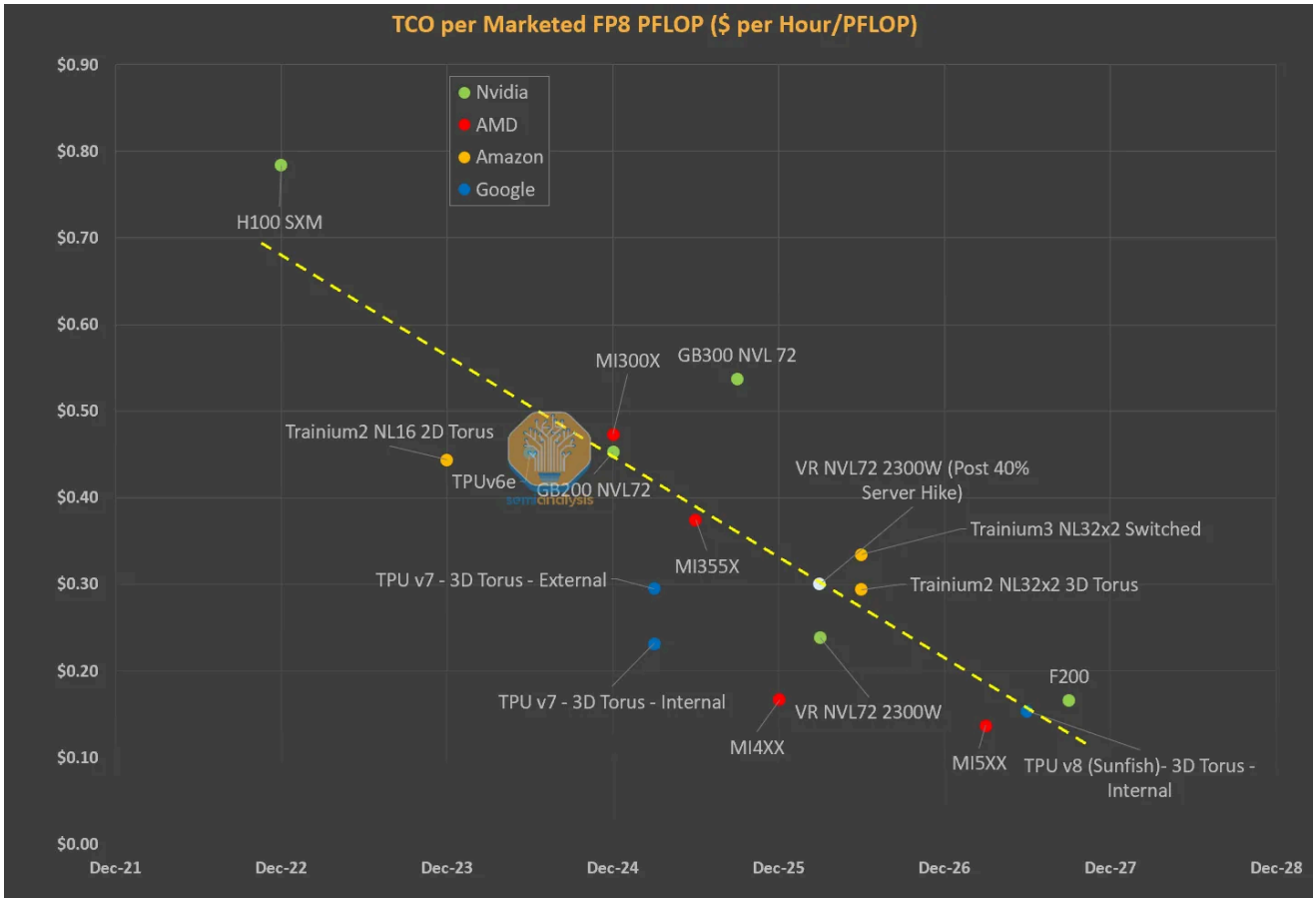
Rubin, set to launch in 2H26, sits at the center of these dynamics. It delivers a step-function improvement in performance but also embeds a much larger memory subsystem at a time when memory is the tightest constraint in the supply chain.

DRAM pricing has already moved sharply higher and is likely to remain elevated, making memory the primary driver of system cost.

In this context, Nvidia has room to increase pricing, particularly for systems like Rubin that deliver step-function performance gains. The incremental value created at the system level far exceeds the incremental cost, especially when viewed through $/FLOP or end workload economics.



Source: AI TCO Model



Source: AI TCO Model

This creates a clear disconnect. The market has structurally shifted, with demand scaling faster and more persistently than supply can respond. Yet Nvidia's pricing framework remains anchored to prior assumptions, rather than adjusting to reflect the increased value its systems now deliver.

Put simply, even if Nvidia raises server pricing and infrastructure providers increase compute pricing, demand would remain intact. Buyers are optimizing for access to compute, end users are optimizing for access to as much tokens as possible, and both are securing capacity at all costs - marginal cost optimization is not their primary concern today.

SOCAMM Pricing: Nvidia's Next Margin Lever

The next question after whether Nvidia can raise prices is where within the system it is most effective to do so.

At the system level, memory is the most natural point of control. Rubin-class systems embed significantly more memory into an already constrained supply chain, and unlike compute, memory can be more cleanly segmented and continuously repriced.

This is because memory on VR NVL72 is a socketed LPDDR-based memory solution called SOCAMM (System-On-Chip Attached Memory Module). SOCAMM is designed for Nvidia's rack-scale systems, enabling higher capacity, modularity, power efficiency, and independent pricing of memory alongside compute.

This makes SOCAMM one of the most important variables in understanding Nvidia's pricing strategy. Two factors ultimately determine system-level pricing outcomes: the cost Nvidia secures for SOCAMM, and the markup applied when reselling that memory to customers. Developing a precise view of Nvidia's pricing and BoM for its rack-scale systems is not an easy task, given the complexity of its "extreme co-design" approach and intricate supply chain dynamics.

This is why SemiAnalysis provides an industry-leading breakdown through our VR NVL72 BoM and Power Budget Model. Furthermore, there are two swing factor at play when determining memory pricing to end customers:

1. The price Nvidia secured for SOCAMM2, and
2. The markup Nvidia applies to SOCAMM when selling to customers,

Both are key factors impacting the final pricing quote to the customers.

	1Q26E	2Q26E	3Q26E	4Q26E
DRAM Pricing by Type ($/GB)				
Mobile DRAM (16GB LPDDR5X)				
Server DRAM (64GB RDIMM)				
PC DRAM (16GB DDR5 DIMM)				
SOCAMM (192GB LPDDR5X Module)				
Blended DDR5 16GB				

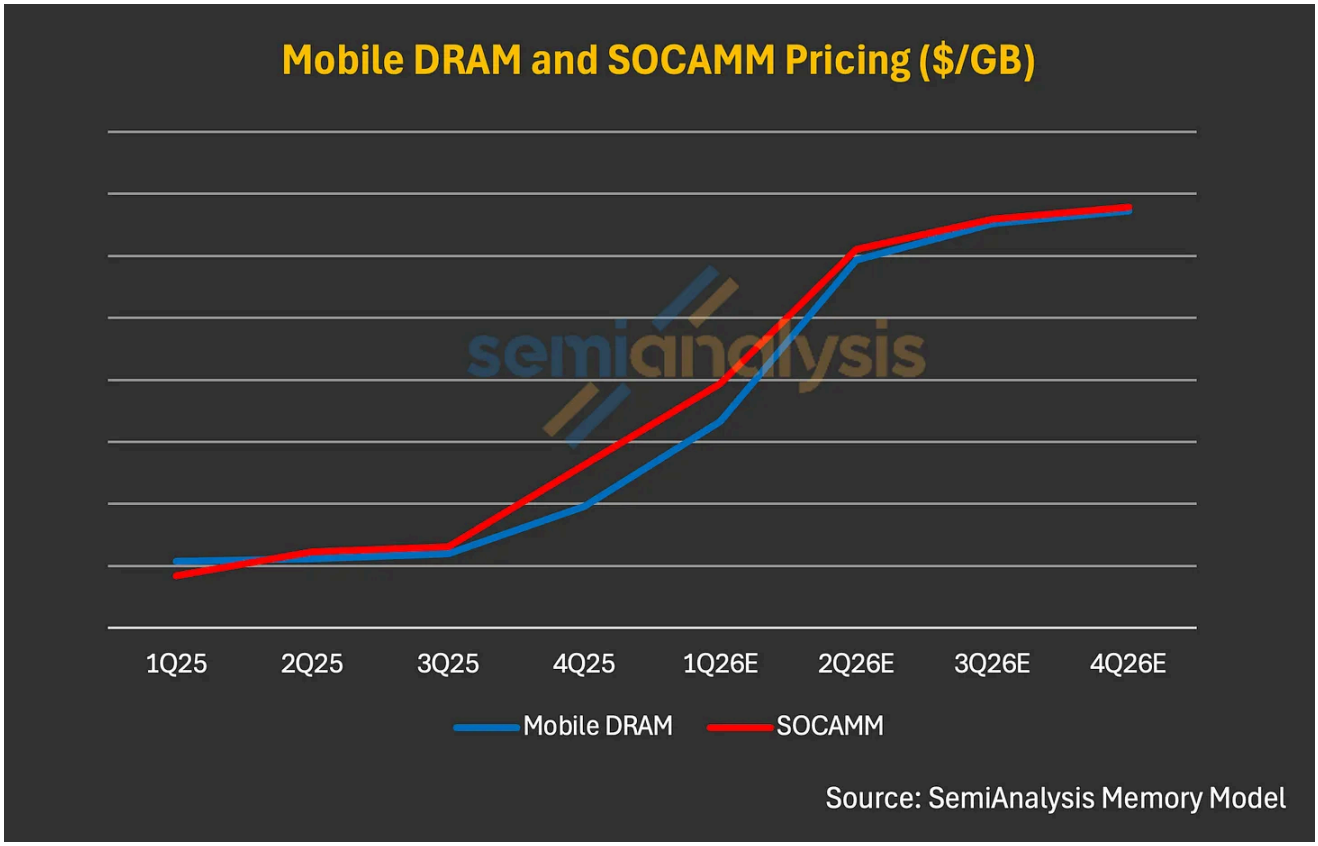
Source: SemiAnalysis Memory Model

As of today, our Memory Model implies SOCAMM contract pricing paid b
~$8/GB in 1Q26, a sharp step-up from 4Q25 to 1Q26. This jump was driven
broader LPDDR5X pricing surge in 1Q and overall memory supply tightne
anchor this estimate based on two points:

1. SOCAMM should price at a premium to mobile LPDDR5X (~$6–7/GB in 1Q26) given higher development complexity and longer cycle times.
2. The step-up in mobile LPDDR5X pricing should transmit to SOCAMM in the same periods, as constrained LPDDR5X and broader commodity DRAM supply is shared between consumer and server demand.

Industry chatter suggests that Nvidia has secured substantial SOCAMM volume for both its GB300 NVL 72 and VR NVL72 systems, under a long-term agreement (LTA) format, which we outlined in our institutional note for Memory Model earlier. As the only scaled SOCAMM customer today, and arguably the most critical buyer in memory, Nvidia likely benefits from preferential access and pricing, and we believe Nvidia's past track record speaks for itself when it comes to its ability to leverage the supply chain.

That said, broader DRAM pricing dynamics should still inevitably flow through. Further price hike in mobile LPDDR5X pricing in coming quarters should still be a critical pricing reference for SOCAMM, and SOCAMM should reprice accordingly given limited LPDDR5 allocation volume. We believe exit '26 pricing for SOCAMM could exceed $13/GB, which is roughly in line with mobile DRAM pricing expected by the end of this year; accordingly, we view ~$10/GB as a reasonable assumption for Nvidia's SOCAMM cost.



Source: SemiAnalysis Memory Model

One key question some may raise is: On what basis should customers accept further price increases and margin expansion from Nvidia, and what rationale can Nvidia credibly use to justify such a position? We think it is reasonable for Nvidia to charge 60% margin on SOCAMM for three reasons:

- First, the current environment plays in Nvidia's hand. Memory supply is constrained everywhere, and Nvidia has secured the most volume (of SOCAMM at least) versus its customers and peer competitors, which should allow the company to leverage this supply chain edge.
- Second, VR NVL72 is still by far the best platform coming to market with regards to performance per TCO, and production of the system backed by a complicated but mature supply chain. To maximize the investment in compute, customers might have little choice but to accept Nvidia's new pricing method.
- Lastly, since Nvidia, as the procurer of SOCAMM2, is facing a material price hike in the first place, we think it is not unreasonable to assume that customers will accept Nvidia's gross margin taken on top of SOCAMM2 cost for VR NVL72.

Capex Per Watt Trends from GB300 to VR NVL72

For GB300, DRAM was bundled into the board and marked up at ~75% gross margin, making the margin charged on the memory on the board consistent with what is implicitly priced for the Blackwell systems.

For Rubin, we initially assumed the same dynamic, with the understanding that Nvidia would target an overall system Gross Margin in the mid-70s. As such, our initial Bill of Material (BoM) modeling applied a consistent margin throughout the entire Strata board leaving SOCAMM margin at the same mid 70s margin.

However, because SOCAMM2 is a socketed module in Rubin whereas GB300 uses an ordinary LPDDR5X module that is soldered onto the board, memory can be disaggregated and quoted separately from the base system. This allows Nvidia to explicitly price memory as its own line item, rather than embedding it within board-level pricing. Importantly, this also introduces an additional value for Nvidia to adjust margin on the SOCAMM2 while keeping margin on the board the same. Even if Nvidia initially absorbs some of the memory cost inflation, it retains the ability to offset this by charging higher system-level margins to customers.

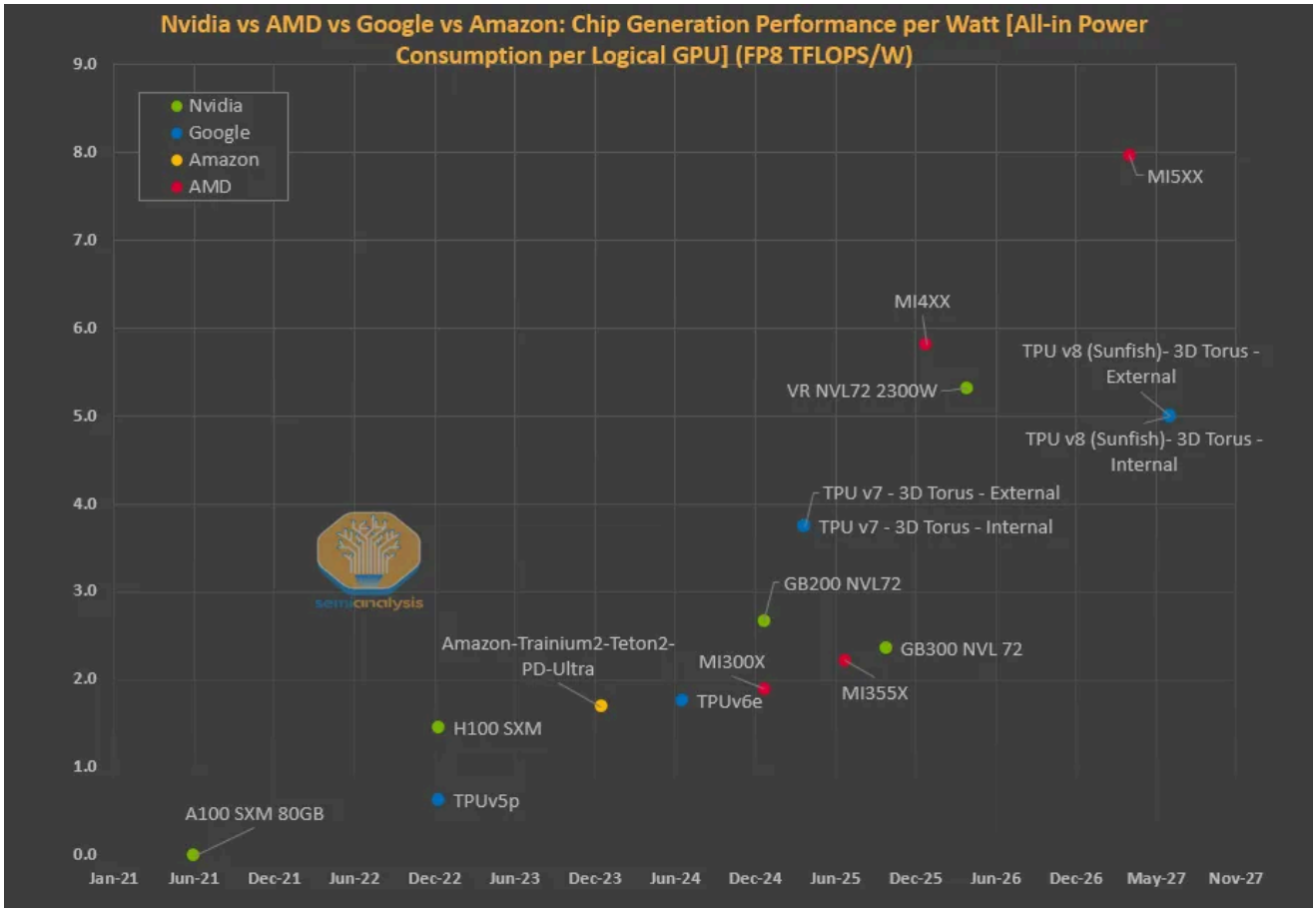
Hence, we would have expected overall capex per watt to rise as we transition from GB300 to VR NVL72. Yet to the contrary – current pricing only works out to a slight creep up in capex per watt from $37.4/W for GB300 to $38.1/W for VRNVL72. This is despite chip TDP almost doubling from GB300 to VR NVL72 (1400W to 2300W), and a material increase in FLOPs.

Capex Breakdown by Power Contribution			
	Unit	GB300 NVL 72	VR NVL72 2300W
Capex per Rack			
GPU Provider Board ASP (Excluding Memory)	$/GPU		
GPU Provider Content	$/Rack		
DRAM	$/Rack		
NAND	$/Rack		
Server Cost	$/Rack		
Networking	$/Rack		
All-in Capex per Rack	**$/Rack**		
Power per Rack			
GPU Provider Board ASP (Excluding Memory)	W/GPU		
GPU Provider Content	W/Rack		
DRAM	W/Rack		
NAND	W/Rack		
Server	W/Rack		
Networking	W/Rack		
All-in Expected Average power, per Rack	**W/Rack**		
Capex Per Watt			
GPU Provider Board ASP (Excluding Memory)	$/GPU		
GPU Provider Content	$/Rack		
DRAM	$/Rack		
NAND	$/Rack		
Server Cost	$/Rack		
Networking	$/Rack		
All-in Capex per Rack per All-in Watt	**$/Rack**	**$37.4**	**$37.6**

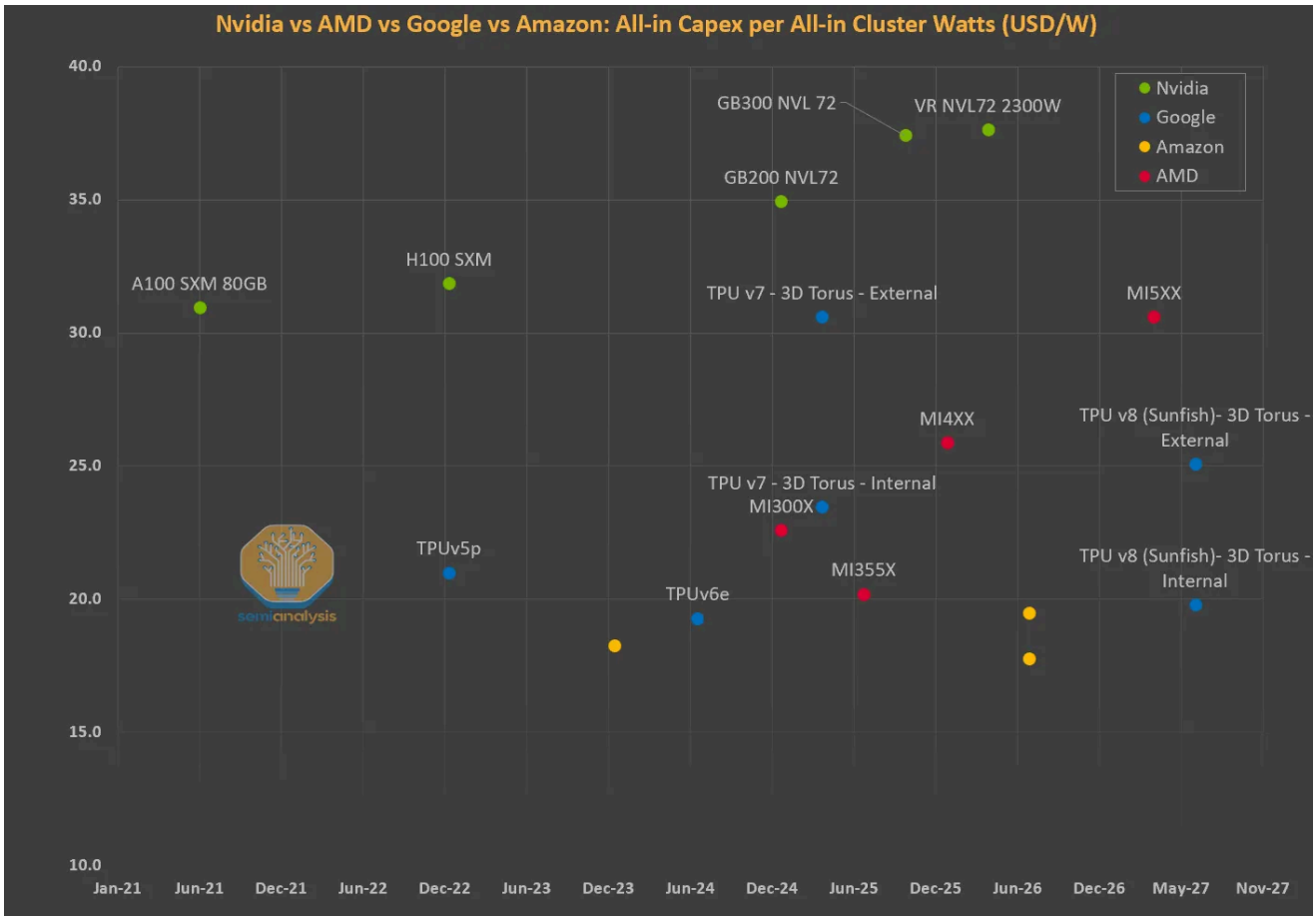
Source: SemiAnalysis AI TCO Model

This is unusual relative to broader trends in server capex per watt. Across AMD, Nvidia, and custom ASICs, capex per watt typically increases generation over generation as improvements in performance per watt allow vendors to capture more value at the system level. Thus, it is puzzling to us that $/GW appears to remain largely stagnant from GB300 to VR NVL72. This is even more unusual given the step up in performance/W from GB300 to VR NVL72 is more than double.

Nvidia also has the opporunity to price discriminate on memory more than they do on the GPU because memory isn't an anti-trust concern whereas the GPU is.



Source: SemiAnalysis AI TCO Model



Source: SemiAnalysis AI TCO Model

Networking as a Vector for Price Discrimination

Today, Nvidia does not heavily differentiate on GPU pricing across customers within a given strata. Core components are generally sold at similar prices across the ecosystem, be it to Hyperscalers, Neoclouds, Emerging Neoclouds, sovereigns or enterprises. This creates a relatively uniform pricing structure for the core GPU and memory components, even in a market where willingness to pay varies significantly.

While GPU pricing is relatively uniform, Nvidia traditionally price discriminates on networking equipment, offering Neoclouds and other marginal cloud players a price point that is at significant premium to hyperscalers. We conducted our own survey with GPU cloud providers and found that, for instance, the SN5610 could be priced 2x more for a Neocloud as opposed to for hyperscalers. Hyperscalers clearly have stronger bargaining power but this is not because they are buying more switches and transceivers from Nvidia.

Neoclouds lack the scale and networking expertise to customize and cost-optimize their networking clusters and so they ultimately prefer Nvidia's turnkey solutions. Hyperscalers work directly with OEMs and ODMs and have the networking engineering bench to deploy more cost effective solutions that may not be turnkey deployments and thus are a heavier lift to deploy properly.

The disparity in networking costs between a Neocloud and hyperscaler becomes less significant, however, on a total cluster capital cost basis. For two comparable clusters, a 94% increase in networking cost for a Neocloud versus a hyperscaler translates to only to a 10% increase in all-in capital cost for a full rack-scale server. This excludes other variables such as power, utilities and operations, which will further erode cost differences attributable to networking equipment pricing disparity.

GB300 NVL72 Cluster Capital Cost Breakdown, per Rack-Scale 72-GPU Server					
	2-Layer 4-Plane Hyperscaler (18,432 GPUs)		2-Layer 4-Plane Neocloud (18,432 GPUs)		%Δ
Item	Cost	%	Cost	%	
Server Cost	[redacted]	84%	[redacted]	76%	0%
Optical Transceivers	[redacted]	4%	[redacted]	8%	140%
Switches	[redacted]	6%	[redacted]	9%	64%
Fiber, Cables, Others	[redacted]	1%	[redacted]	2%	96%
Networking Cost	[redacted]	11%	[redacted]	19%	94%
All Others	[redacted]	5%	[redacted]	5%	0%
All-in Cost per Rack-Scale Server	[redacted]	**100%**	[redacted]	**100%**	**10%**
Costs for a Spectrum-X cluster using Nvidia SN5610 Back-end switches with 64 ports of 800G					

Source: SemiAnalysis AI Networking Model

Though this is a great case study demonstrating how Nvidia is pricing its solutions to value, the current price gap between Neoclouds and Hyperscalers is meaningful, leaving limited room for Nvidia to pull this lever further.

Nvidia as the Central Bank of AI

One explanation behind Nvidia restraint's in pricing thus far might be a combination of regulatory and strategic reticence.

Nvidia's position in the AI compute stack is already under increasing antitrust scrutiny, given its dominance across GPUs, interconnect, and software. In this environment, aggressively repricing systems to fully capture the value delivered risks drawing further attention, particularly if it results in outsized margin expansion while downstream AI labs are also generating significant profits. Holding pricing closer to prior frameworks can help avoid signaling excessive pricing power in a supply-constrained market.

This behavior is not without precedent. TSMC has historically taken a similar approach. Even while operating at full utilization and acting as the bottleneck for advanced-node supply, TSMC has generally avoided fully pricing to scarcity. Instead, it has prioritized long-term relationships and ecosystem stability over extracting maximum short-term margins, in part to avoid regulatory and customer backlash.

Nvidia appears to be following a comparable path. Rather than fully repricing Rubin systems to reflect both the increase in performance and the structural shift in memory

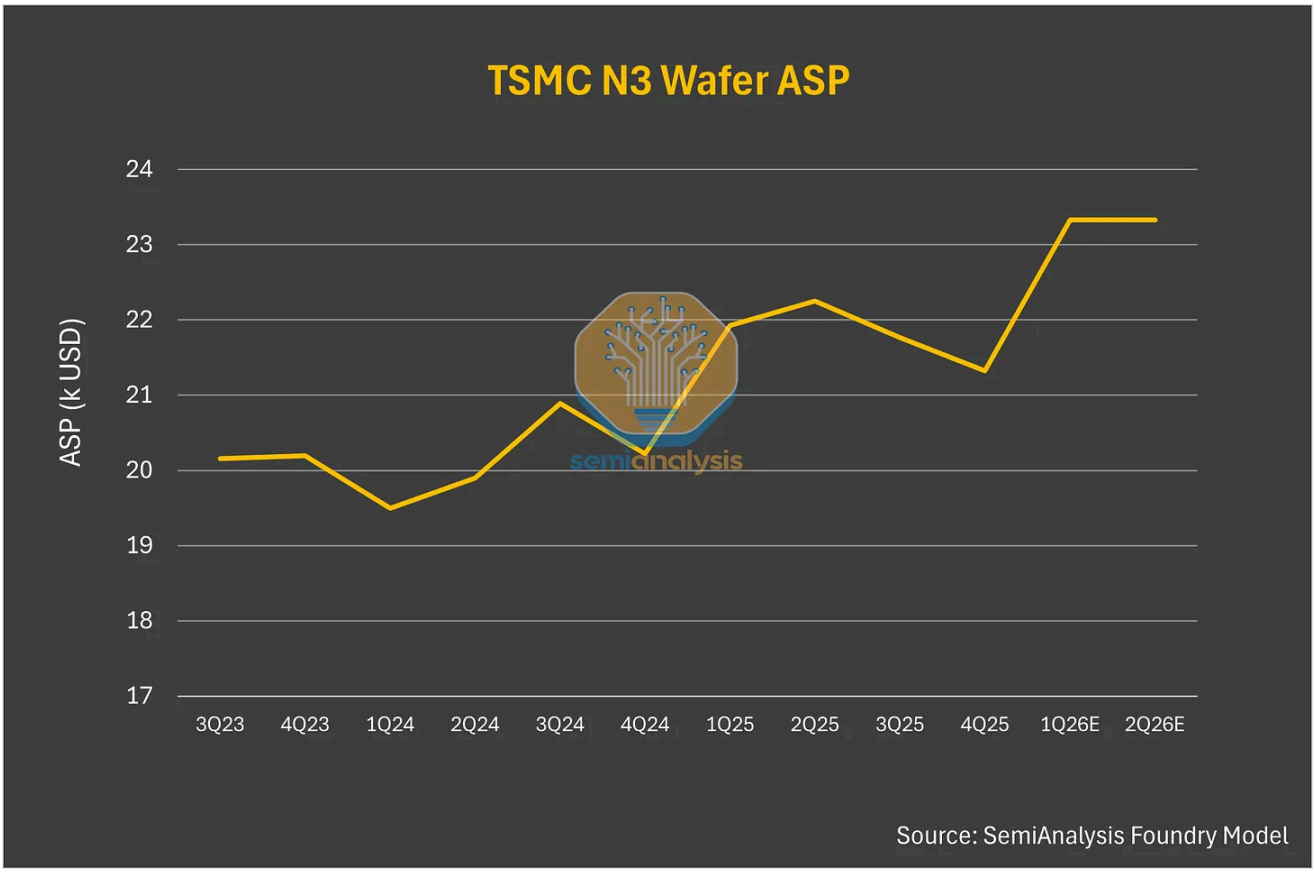
costs, it is maintaining a more measured pricing approach. This balances margin expansion against regulatory risk, ecosystem dynamics, and the need to avoid accelerating customer diversification toward alternative compute platforms.

We made a similar point in our Nvidia as the central bank note. Nvidia is actively supporting the development of the broader ecosystem, ensuring long-term demand expansion rather than maximizing near-term extraction. Today, frontier labs benefit from Nvidia's software-driven efficiency gains, but these improvements are not fully monetized at the hardware level. As a result, incremental value continues to accrue downstream despite Nvidia being the primary enabler. By taking the oxygen out of the room – Nvidia aims to ensure it remains the main protagonist in the AI era for the foreseeable future.

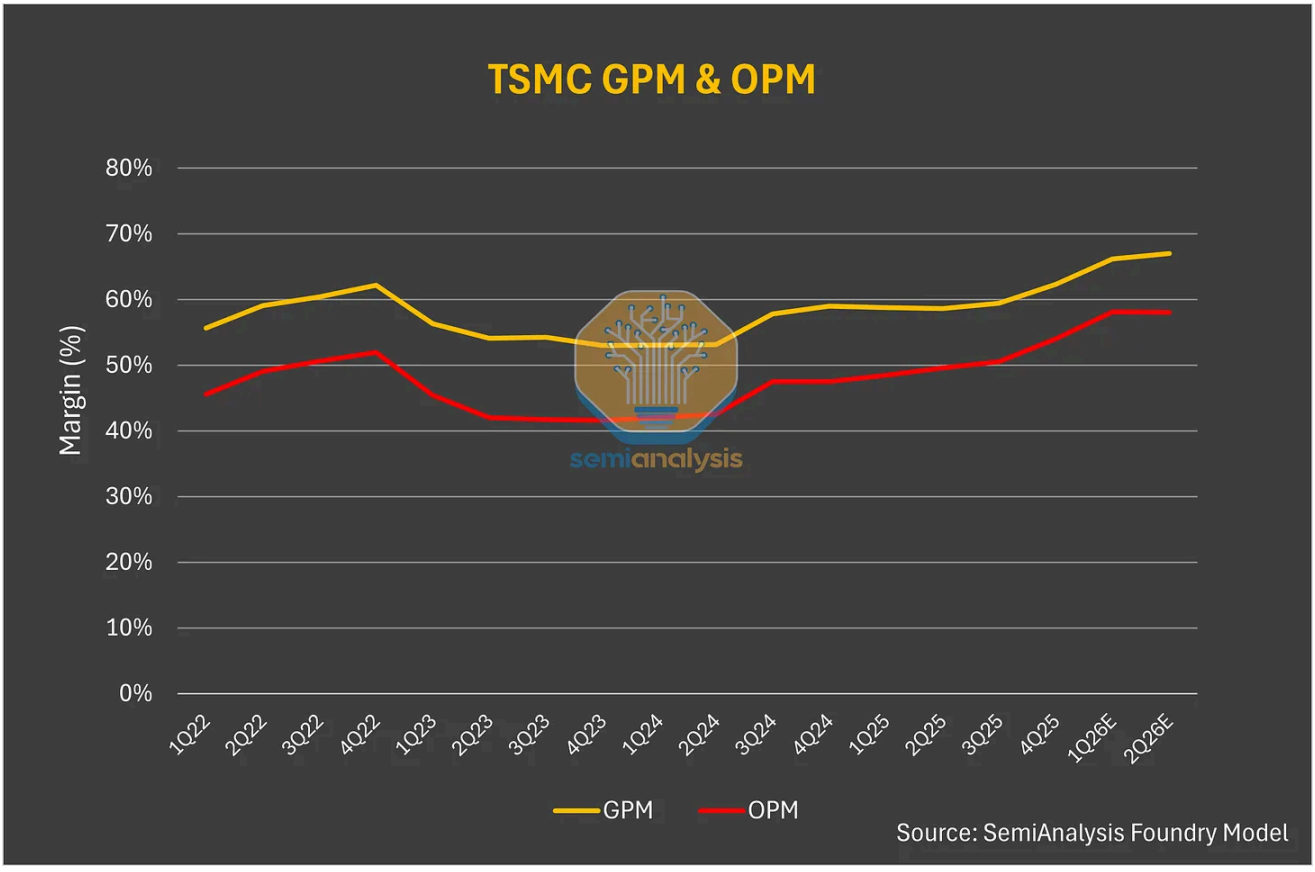
Yet - with Compute demand well over compute supply, why should those in control of the scarce resources not capture higher pricing and enjoy greater profitability?

TSMC, The Fairest and Most Just Company In The World

We've said that TSMC's N3 capacity is even tighter. All major accelerator roadmaps have now converged on the N3 process node for this year and next year. Nvidia, Broadcom, Annapurna, MeidaTek and AMD are all fighting for more N3 wafer allocation from TSMC so that they can ship more compute to their customers. While N3 capacity is arguably the tightest constraint in the system, pricing remains relatively stable.



Source: SemiAnalysis Foundry Industry Model



Source: SemiAnalysis Foundry Industry Model

TSMC strategically looks to protect profitability through downcycles. The flipside is that this policy also blunts upside during upcycles. Regardless, TSMC is certainly leaving value on the table with all their major fabless customers enjoying very high gross margins that could be transferred over instead to TSMC.

However, TSMC could very much take a more aggressive posture on pricing, and customers would not only accept it but we would argue some would even welcome it. Nvidia would love to pay more for wafers if it means shutting out their competition who have less ability to pay up. After all, Jensen himself said in 2024 that TSMC should charge more for wafers and he meant it for this reason.

TSMC can also take a position of longer term agreements with guaranteed capacity commitments and prepayments in liu of major price increases. This is the more likely path.

We think Nvidia is starting to look a lot like TSMC.

Its greatest strength in this environment is procurement. Nvidia has secured disproportionate access to constrained upstream supply, particularly TSMC wafers, allowing it to serve demand that others cannot.

AI compute buyers such as Anthropic are therefore forced into Nvidia's ecosystem, as alternative capacity from TPU and Trainium remains limited by the same upstream bottlenecks. Despite this structural advantage, Nvidia is not fully reflecting it in pricing.

For now, pricing for Nvidia remains anchored to cost-based frameworks. But this is unlikely to hold. As the return on investment for inference providers becomes clearer and more widely accepted, the focus will shift even more towards pricing to value. This reduces the scrutiny on pricing and gives GPU infrastructure providers room to move from cost-based to value-based pricing. Once that transition occurs, it creates space for Nvidia to move pricing higher and capture more of the value delivered at the system level – the manifestation of the pie growing.

Triangulating VR NVL72 Rental Pricing: Cost-Based vs Value-Based Approaches

There are two main approaches to pricing:

1. Cost-based Pricing and,
2. Value Based Pricing.

The Cost-based pricing approach starts with the premise that GPU deployments will only occur if these projects they meet a minimum return threshold for Neoclouds. If returns fall below this level, capacity will not be deployed until pricing adjusts to meet that hurdle.

Therefore, the rental price charged under a cost-based framework is the price that earns the Neocloud a project IRR above the minimum hurdle rate for deployment. Most projects today tend to earn a return of mid-to high teens IRR. An illustrative GB300 deployment today will likely have a project IRR of 15.6% over a 5-year period with a 15% prepay.

Scenario Settings

System Configuration	**GB300 NVL 72**	[redacted]
Number of Accelerators in Project	72	
Price per Server	[redacted]	**USD**
Logical GPUs per Server	72	
Number of Servers	[redacted]	
Total Server Capex		**USD**
System First Production Date	**30-Sep-25**	
Cost of Capital for NPV	[redacted]	
Customer Prepay %		
Customer Prepay Period Basis		Years
Locked in term		Years
Customer Locked-in Rental		USD/hr/Chip
Customer Prepay Amount		USD
Physical Chip Expected Lifetime		Years
Shut down EBITDA Margin		%
	EQ IRR	8.0%
Costs	Proj IRR	15.6%

Source: SemiAnalysis AI TCO Model

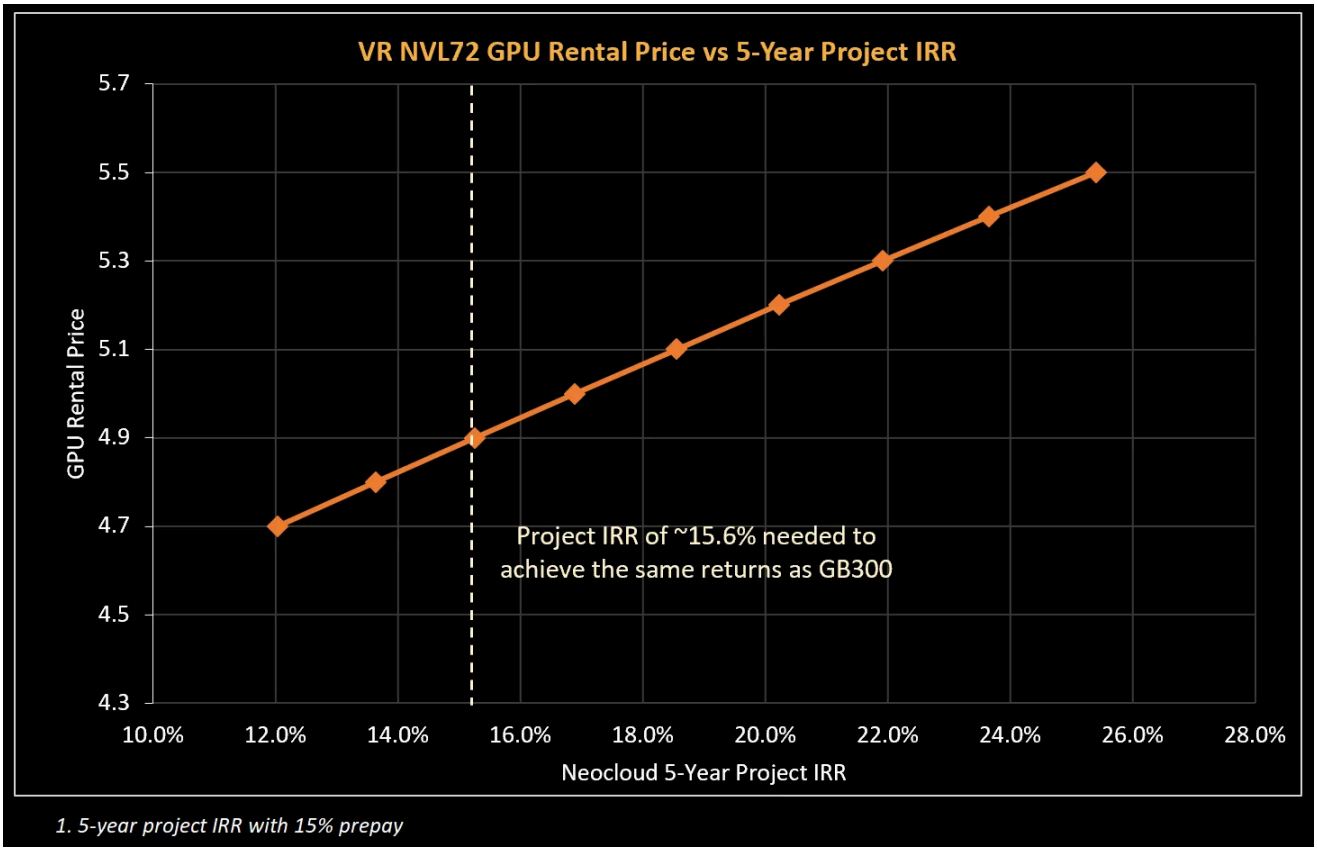
Neoclouds will aim for a similar IRR when deploying VR NVL72, which in turn determines what debut GPU rental prices for Vera Rubin might look like. With our all-in server cost for VR NVL72, a rental price of at least USD 4.92 per Hour per GPU rental price is required for a 5-year project with a 15% prepay to achieve the same project IRR hurdle of 15.6% that most GB300 projects use.

Project Assumptions and Setup

Scenario Settings

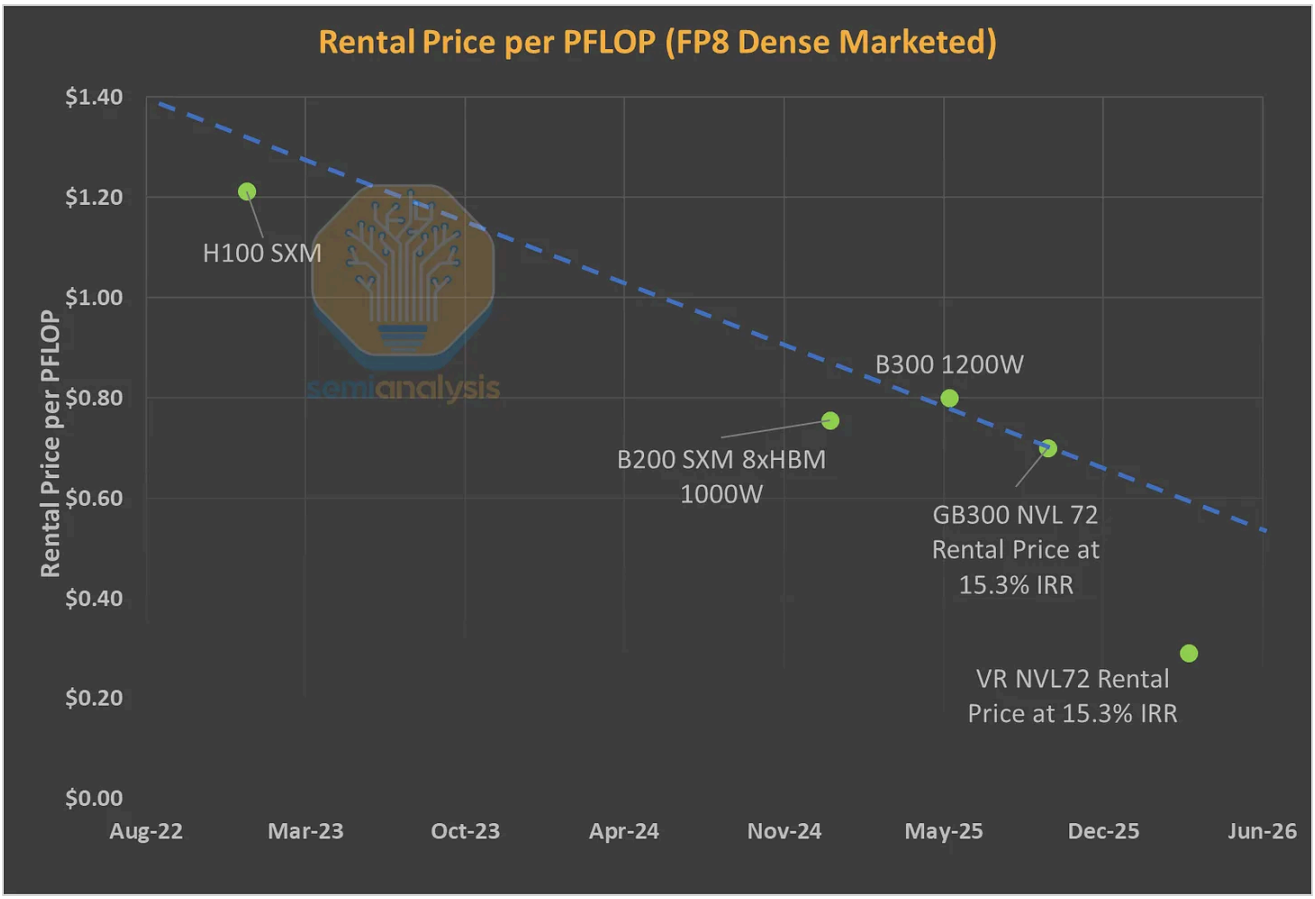
System Configuration	VR NVL72 2300W	[redacted]
Number of Accelerators in Project	72	
Price per Server	[redacted]	**USD**
Logical GPUs per Server	72	
Number of Servers	[redacted]	
Total Server Capex		**USD**
System First Production Date	**31-Mar-26**	
Cost of Capital for NPV	[redacted]	
Customer Prepay %		
Customer Prepay Period Basis		Years
Locked in term		Years
Customer Locked-in Rental	$4.92	USD/hr/Chip
Customer Prepay Amount	[redacted]	USD
Physical Chip Expected Lifetime		Years
Shut down EBITDA Margin		%
	EQ IRR	8.1%
Costs	Proj IRR	15.6%

Source: SemiAnalysis AI TCO Model



Source: SemiAnalysis AI TCO Model

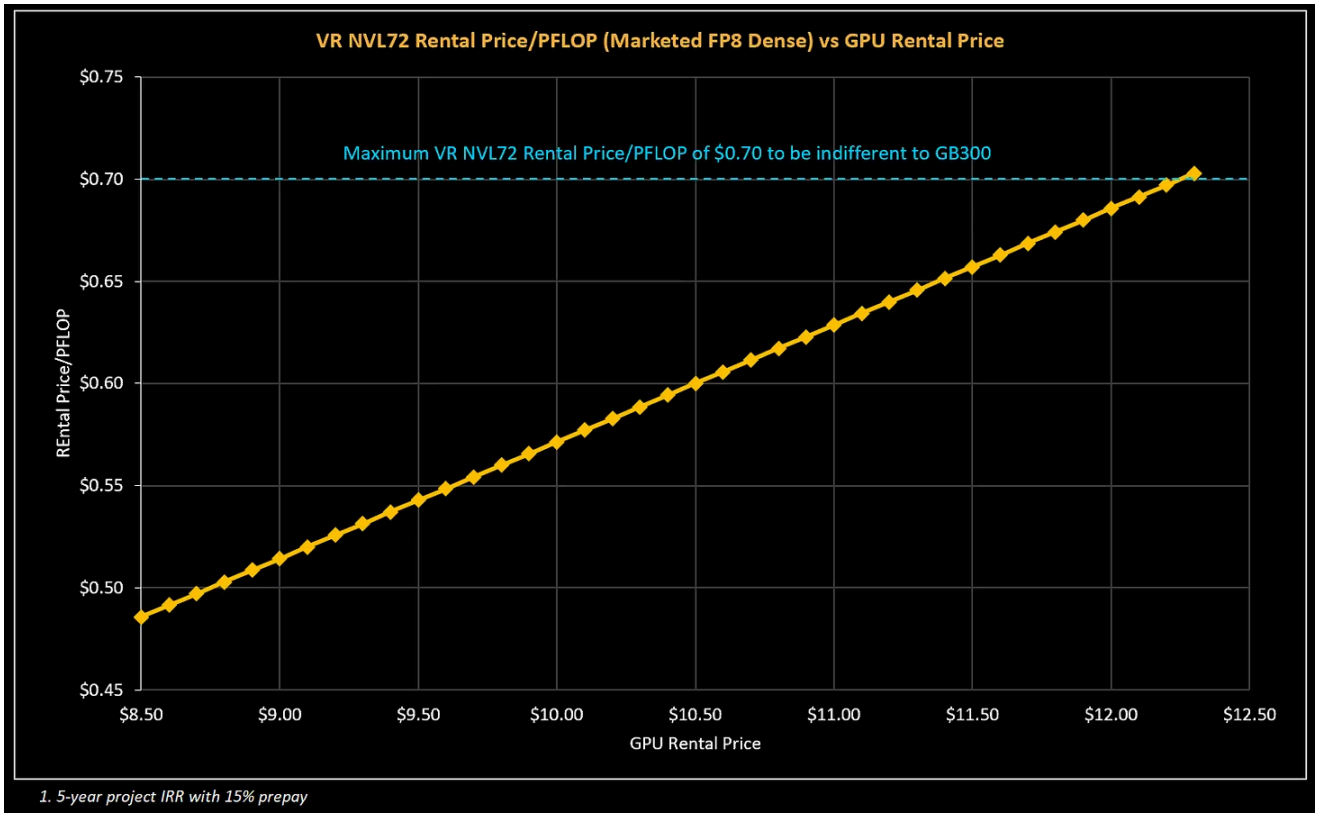
Our second framework looks at value-based pricing. We anchor on the $/FLOP implied by existing SKUs and ask what that would translate to for Rubin. This represents the theoretical maximum a renter of compute would be willing to pay to remain indifferent between Rubin and current generation GPUs - and therefore serves as the ceiling for GPU rental pricing. Here - we look to the trend in improvement of rental prices per PFLOP.



Source: SemiAnalysis AI TCO Model

For training workloads, we anchor on GB300 pricing by comparing rental cost per PFLOP on a marketed FP8 dense basis. Using a current 5-year GB300 rental price of ~$0.70 per PFLOP, we derive a VR NVL72 ceiling price of approximately $12.25 per GPU hour at parity.

The VR NVL72 price per TCO stands out in that, unlike for the GB300 and prior cards, there is an extremely large gap in Value-based and Cost-based pricing. If we are conservative and select a point slightly below the trend line - for instance a rental price of $0.55 per PFLOP - this would corresponds to $9.63/hr/GPU, nearly double the minimum rental price of $4.92/hr/GPU needed to cross Neoclouds' return hurdle.



Source: SemiAnalysis AI TCO Model

One Chart to Rule Them All

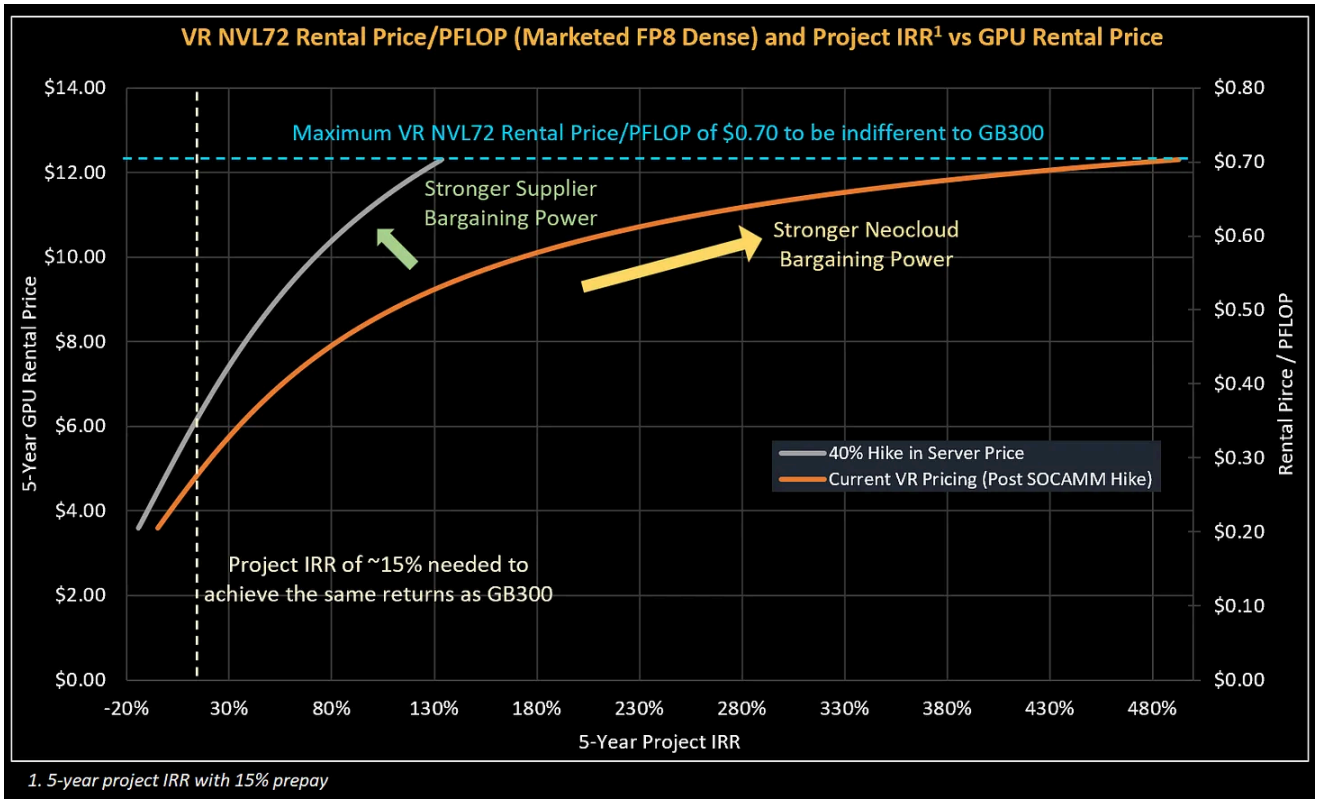
The cost-based approach forms the floor for GPU rental pricing – below this rental price, Neoclouds will not greenlight new GPU projects. The value-based approach forms the theoretical ceiling for GPU rental pricing – no customer would pay higher on a $/FLOP basis to rent a newer generation GPU.

We combine both constraints as well as a pricing curve that illustrates the returns to the Neocloud for given GPU rental prices to create one pricing chart. This “One Chart To Rule Them All” also acts as a framework to understand competitive dynamics and pricing power.

At the start of the article – we posed the question: Who are the benefits of strong AI demand accruing to?

This question can be answered by plotting observed GPU Rental Prices charged by Neoclouds and the IRRs earned by these projects. Sliding up and to the right along the orange curve in the chart below represents stronger Neocloud bargaining power: The Neocloud is able to charge higher GPU Rentals and earn well above their IRR hurdle rate.

If Nvidia increases pricing for VR NVL72, the pricing curve shifts up and to the left. This is because a higher rental price is needed to offset this higher system cost while still earning the same IRR from the Neocloud perspective. This shift represents stronger bargaining power enjoyed by system suppliers like Nvidia.



Source: SemiAnalysis AI TCO Model

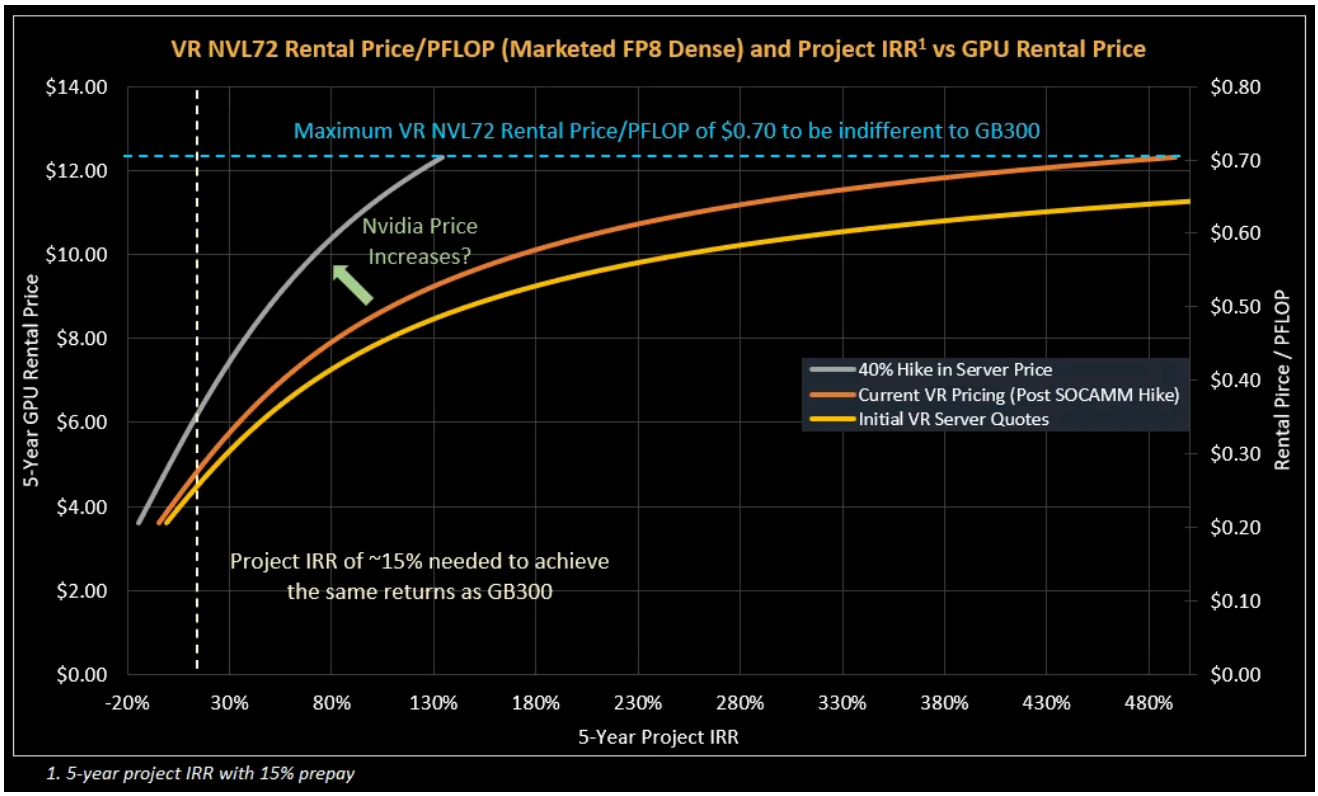
The top left corner, where the blue maximum rental price/PFLOP and the beige Neocloud Project IRR minimum hurdle intersect, represents the maximum theoretical AI Cluster pricing. If the system is priced any higher, Neoclouds and end users would be better off just purchasing or renting GB300s. The larger the gap between the current pricing curve and the top left corner, the more room there is for AI Cluster providers like Nvidia to increase system pricing.

At today's VR NVL72 system pricing, Neoclouds can charge $4.90/hr/GPU for a 5-year contract while still earning the same 15% IRR as they do on their GB300 projects. For customers – Rental Price per PFLOP works out to $0.28/PFLOP, a 60% drop in cost per PFLOP vs the GB300 NVL72, an improvement in cost that is well below trend.

This suggests that there is meaningful room for Nvidia to increase server prices. A ~40% increase in server pricing would deliver below trend cost improvements in price per FLOP, while still leaving Neoclouds enough room to lift prices even higher so they

can earn higher IRRs. Even if Neoclouds adjust pricing higher to slide along the grey curve, for instance charging $8.00/hr/GPU and earning a 38% IRR, corresponding to a cost of $0.46/PFLOP, which is still an improvement that is below trend.

It is important to point out that this analysis has mainly focused on Rental Price/FLOP - but improvements in inference performance per TCO have been accelerating at an even brisker pace. Though we have yet to benchmark a VR NVL72 system at InferenceX, it is highly likely that there is an even sharper pace in cost decreases when it comes to dollars per token delivered by VR NVL72, meaning there could be even more headroom for Nvidia to capture more value from the overall ecosystem.



Source: SemiAnalysis AI TCO Model

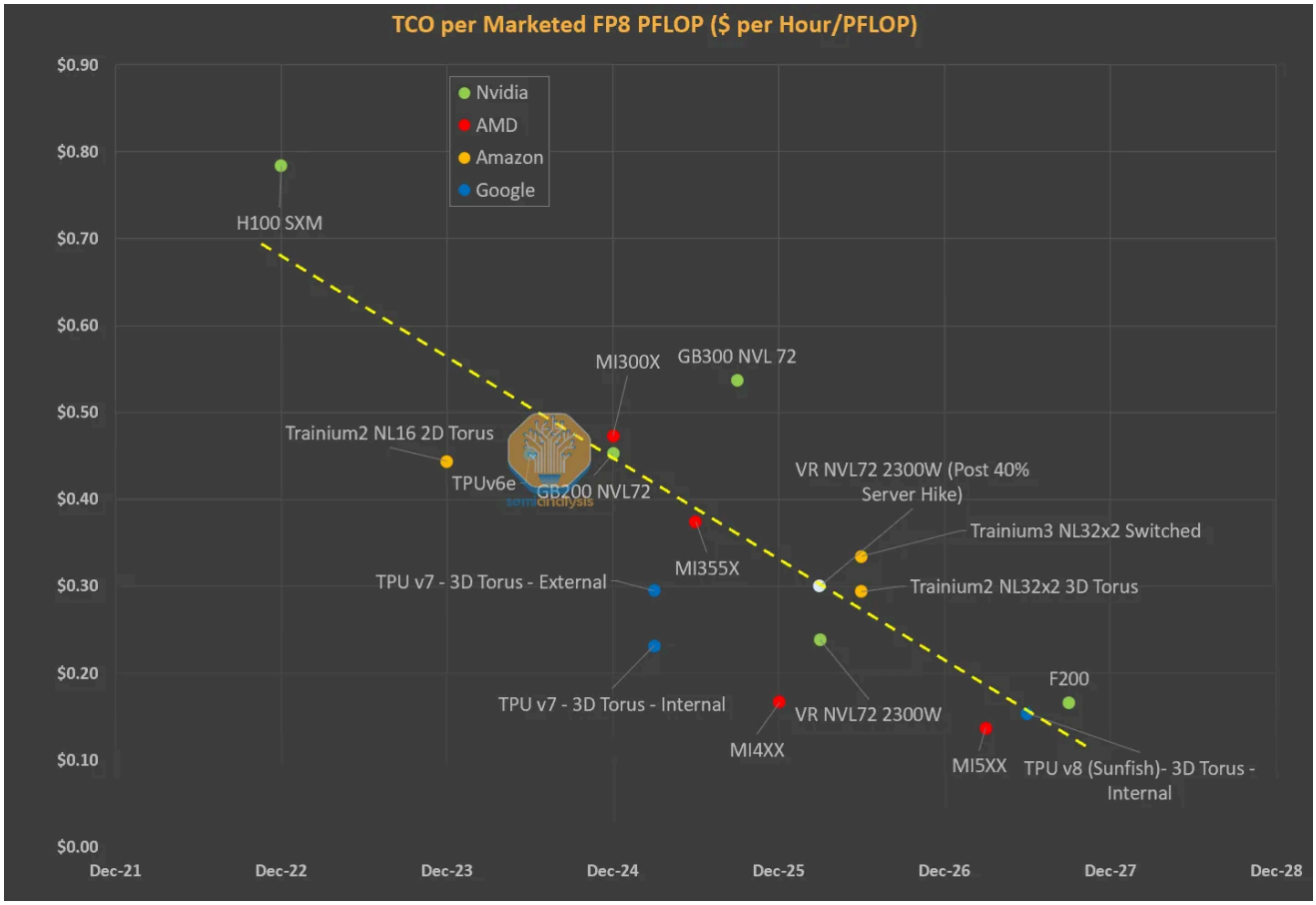
VR NVL72 vs GB300 Performance per TCO

Rubin delivers a clear step-up in absolute performance, but the more important question is how that translates into performance per TCO.

At the system level, VR NVL72 carries a higher cost base. Total cost per GPU per hour increases from $2.69 for GB300 NVL72 to $4.18 for VR NVL72, reflecting a higher system cost that is driven in no small part by memory price hikes.

Marketed dense BF16 performance increases from 2,500 TFLOPS to 4,000 TFLOPS, while Marketed dense FP8 scales from 5,000 to 17,500 TFLOPS. The largest jump comes from FP4, where dense performance rises from 15,000 to 35,000 TFLOPS. Memory bandwidth also increases significantly, from 8 TB/s to 22 TB/s per logical GPU. Nvidia is also marketing FP4 sparsity, quoting 50,000 TFLOPS versus 35,000 TFLOPS dense. If this can be effectively utilized, it drives a further reduction in cost per performance, with TCO per PFLOP improving from $0.12 to $0.08, a ~33% reduction. This is a meaningful lever, but it is conditional on real-world workload compatibility.

On a TCO basis, Rubin shows modest improvement for BF16, with cost per PFLOP declining slightly from $1.07 to $1.04 per hour. The improvement becomes more pronounced at lower precision. FP8 TCO per PFLOP drops from $0.54 to $0.24, while FP4 dense improves from $0.18 to $0.12. These gains reflect the architectural shift toward lower precision compute and Nvidia expectation that workloads will increasingly move down the precision stack.



Source: SemiAnalysis AI TCO Model

These comparisons are based on marketed performance rather than effective throughput. Real-world performance depends on model FLOPs utilization (MFU), which varies significantly by workload and familiarity with the system, which is also another gating factor for Rubin to reach value-based pricing when it's first deployed, given MFU% will likely be lower on initial deployment before the software support and engineering know-how matures.

One important change versus our prior February newsletter article is how BF16 performance scales relative to FP8 and FP4. Our earlier assumption was that all three precisions would scale together, driven by a uniform increase in flops per clock at the Streaming Multiprocessor (SM) level. The final Rubin specifications show a different picture. BF16 flops per clock per SM remain largely unchanged from Blackwell, while FP8 and FP4 see a doubling in flops per clock per SM. As a result, BF16 performance only increases modestly, driven primarily by higher SM counts and incremental clock improvements, rather than any architectural uplift at the tensor core level.

This shifts the performance mix meaningfully toward lower precision, reinforcing that the bulk of Rubin's performance gains, and therefore its TCO advantages, are concentrated in FP8 and FP4 rather than BF16.

AI Cloud Operating Cost of Ownership Summary			
Chip	**Unit**	**GB200 NVL72 (Spectrum)**	**VR NVL72 2300W (Spectrum)**
Customer Profile	**Unit**	**Neocloud Giant**	**Neocloud Giant**
Total Cost per Unit per Hour	**USD/hr/GPU**	**$2.26**	**$4.22**
Capital Cost as % of Total Ownership Cost	%	73.9%	76.0%
Marketed TFLOPS (FP8)	TFLOPS	5,000	17,500
Effective training TFLOPS (FP8)	TFLOPS	2,500	7,875
Inference Throughput[1]	Tok/s/GPU	8,850	24,579
Memory Bandwidth per Logical GPU	TB/s	8	22
Marketed TFLOPS (FP8) / Memory Bandwidth	TFLOPS/TB/s	626	788
TCO per PFLOP	$/hr per PFLOP	$0.45	$0.24
TCO per effective training PFLOP	$/hr per PFLOP	$0.91	$0.54
TCO per M Tokens	$/M tokens	$0.07	$0.05
TCO per Memory Bandwidth	$/hr per TB/s	$0.28	$0.19

1. DeepSeek R1 FP4. Uses 8k input, 1k output and 1k input, 1k output tokens 50/50 mix, 100 Interactivity.

Source: SemiAnalysis AI TCO Model

Compute Competition

Profitability is ultimately dictated by competition. When competition is intense, pricing converges closer to cost (i.e. lower margins), while monopolies can price to value or the maximum amount that customers can bear (higher margins).

Nvidia is supplying scarce compute into a market where customers like Anthropic are generating strong inference margins, indicating a significantly higher willingness to pay for compute than current pricing reflects.

However, this is ignoring Nvidia's competition at the compute level. One of the key stories last year was how Anthropic has been able to successfully diversify their compute away from Nvidia systems. Mythos was not trained on Nvidia. Nvidia has long been the incumbent merchant GPU provider of choice for just about every company except for Google.

First, Anthropic pivoted heavily towards Trainium and TPU, a development that we covered in detail in an earlier article. While TPU and Trainium do not have absolute hardware and software superiority to Nvidia GPUs, this shortcoming is made up for by lower cost. Amazon and Google pay lower margins to their design partners like Marvell, Alchip, Broadcom, and Mediatek than Nvidia charges their customers, and are still able to resell compute to bring down cost per token or cost per training FLOP. This is the competitive gravity that some may argue is keeping Nvidia margins from moving up.

Even if Nvidia maintains a clear performance lead, the presence of credible lower-cost alternatives limits how aggressively it can move pricing without accelerating customer diversification.

Ultimately, the question is whether Nvidia is able and willing translate its performance advantage into pricing power.

Our verdict is they should strike while the iron is hot and take advantage of their long term advantages in memory pricing, capacity, and performance. d

---

Recommend SemiAnalysis to your readers

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

Recommend

71 Likes · 8 Restacks

← Previous

A guest post by

Crystal Huang

Subscribe to Crystal

Discussion about this post

Comments **Restacks**

Write a comment...