AI Value Capture - The Shift To Model Labs

AI 价值捕获——向模型实验室的转变

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?

Vera Rubin VR NVL72：V 代表价值 —— Rubin 在单位总拥有成本（TCO）性能上实现了跨越式提升。投资回报率（ROI）将流向用户、新型云服务商（Neoclouds）、超大规模云厂商（Hyperscalers）、AI 实验室、存储芯片厂商还是 GPU 制造商?

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

DANIEL NISHBALL, DYLAN PATEL, CHEANG KANG WEN, 以及其他 6 位作者

MAY 01, 2026 2026 年 5 月 1 日 · PAID · 付费内容

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.

AI 领域的一天现在感觉就像其他行业的的一年。模型发布、软件突破和硬件改进正将其他行业长达数年的周期压缩至数周。就在过去的几个月里，智能体 AI（Agentic AI）已经跨越了一个真正的拐点，推动了 Token 价值的阶梯式增长，而软件和硬件的改进则大幅降低了生成 Token 的成本。

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.

这种需求的激增是由终端用户从消耗 Token 中获得巨大的投资回报率（ROI）所驱动的，而且这种需求的增长可以说还处于早期阶段。今年，Anthropic 的年度经常性收入（ARR）已从 90 亿美元爆发式增长到如今的 440 亿美元以上，同期其推理基础设施的毛利率也从 38% 提升至 70% 以上。

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.

这种快速的 AI 普及进程已在整个技术栈中创造了价值，但一个独特的现象是，AI 实验室目前正占据着全部价值，而去年这一比例几乎为零。

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.

最终用户正享受着一场生产力盛宴，过去需要数十个人工时、耗资数千美元的任务，现在只需价值几美元的 token 即可在几分钟内完成。这种收入和利润率的巨大增长，是因为所生成的 token 价值正在显著改善业务。例如，SemiAnalysis 在 Anthropic Claude token 上的年度支出率已高达 1095 万美元，但我们从中获得的价值使我们能够超越所有竞争对手并赢得市场份额。

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.

与一年前的 Hopper 相比，Blackwell 等新芯片在运行当今的前沿工作负载时，每秒生成的 token 数量可提高 30 倍，而 TPUv7 和 Trainium 3 等 ASIC 也显示出类似的改进。Fireworks、Baseten、Fal 等推理服务提供商的利润率正在扩大，同时其营收趋势正处于高速增长中。

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.

甚至硬件堆栈的部分环节也经历了重新定价，内存价格在过去一年中上涨了 6 倍。Neocloud GPU 租赁价格也在飙升，1 年期 H100 租赁合同价格较 2025 年 10 月的低点上涨了 40%。

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.

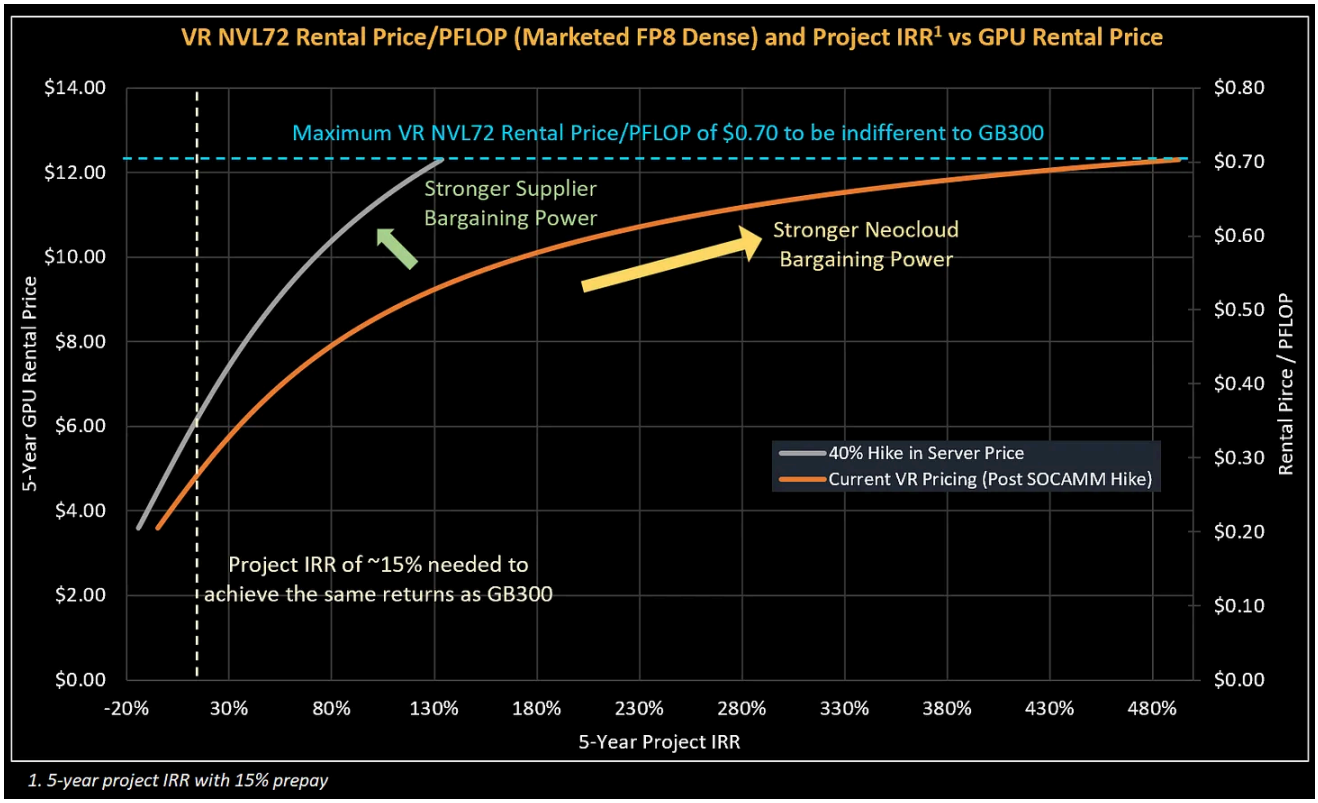
该行业中有两家拥有惊人定价能力的公司，但其表现却并未出现太大波动。台积电（TSMC）和英伟达（Nvidia）尚未对近期 AI 模型价值生成的爆发式增长做出反应。

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.

在这篇文章中，我们将探讨 AI 产生的价值正流向何方——从终端用户到推理提供商、Neoclouds 以及硬件供应商。我们将揭示 TSMC 和 Nvidia 是如何将巨大的价值输送到生态系统的每一个垂直领域的。

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.

最后，我们引入了一个全新的框架："统领全局的图表"（One Chart to Rule Them All），旨在探讨 GPU 租赁经济学，并分析在终端用户、Neoclouds/超大规模云厂商（Hyperscalers）以及 AI 系统供应商中，谁正在捕获 AI 生态系统中最大的价值。



Source: SemiAnalysis AI TCO Model

SemiAnalysis AI TCO 模型

AI Value Profit Pools AI 价值利润池

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.

从 2023 年到 2025 年，AI 领域的所有价值都被基础设施层所捕获。Nvidia 在 2023 年 5 月发布了首份轰动性的财报，股价在盘后飙升 25%，正式开启了 AI 交易浪潮。2024 年，随着人们意识到电力正成为关键瓶颈，Vistra 和 GE Vernova 成为标准普尔 500 指数中表现最好的两只股票（分别上涨 265% 和 146%）。2025 年，存储芯片大放异彩，SanDisk、Western Digital、Seagate 和 Micron 全年涨幅均超过 200%。当然，这些都是概括性的总结，由于 AI 资本支出的增加，许多其他基础设施公司的表现也远超大盘。对所有细分详情感兴趣的读者，请订阅我们的机构产品。

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.

在此期间，所有模型创作者和推理提供商的毛利率之低是众所周知的。对大多数人而言，AI 的实际效用仍然仅相当于锁定在聊天界面背后的、略好一点的 Google 搜索，以及吉卜力工作室风格的自拍。怀疑论者大声宣称，AI 根本不可能兑现数万亿美元的计划资本支出。

Agentic AI Has Changed the Game

智能体 AI 已改变游戏规则

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.

2025 年 12 月，世界发生了改变，代理式 AI（Agentic AI）开始真正发挥作用。SemiAnalysis 曾多次撰文并深入探讨我们对 Claude Code 的使用情况，但必须强调的是，代理式 AI 的应用早已不再局限于编程。我们的分析师每天都在使用智能体将 Excel 模型转换为仪表板、为所有笔记创建图表、构建财务模型并分析公司收益等等。这些任务要么是 1) 我们以前根本无法完成的，要么是 2) 以前需要初级分析师花费数小时才能完成，从而让他们无法专注于更有价值的任务。

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:

下表展示了我们自身工作流中的几个真实案例，对比了 Token 支出与等效人力成本的费用：

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 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.

SemiAnalysis 的年化 Token 支出已达到员工薪酬的约 30%，我们每位员工每月消耗的 Token 数量略低于 50 亿个（是 Meta 的 5 倍多！）。不过，这一分布遵循幂律定律，因此有些团队成员每月的消耗量超过 1000 亿个 Token。显而易见，这仅仅是个开始，所有白领企业很快都将拥抱智能体 AI（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.

在过去的几个月里，每个 token 的价值显然有所提升。我们估计，在执行智能体（agentic）任务时，运行 Opus 4.7 的实际混合价格为每百万 token 0.99 美元，尽管其标价为每百万 token 5 美元/25 美元。智能体工作负载具有极高的输入输出比（我们的 Claude Code 使用比例约为 300:1）和高缓存命中率（90% 以上）。由于缓存的输入 token 成本仅为 0.50 美元/百万 token，大部分 token 最终都落入了最便宜的价格档位。我们在这里详细介绍了完整的计算方法。

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

从这个角度来看，就不难理解为什么 Anthropic 的 ARR（年度经常性收入）在今年以来从 90 亿美元爆发式增长到可能超过 440 亿美元了。

Tokens Are Getting Cheaper to Produce

Token 的生产成本正在降低

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.

与此同时，生成每个 token 的成本已大幅下降。这是推理服务提供商价值增长的最大驱动力，也是大型 AI 实验室利润率大幅提升的关键原因。

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.

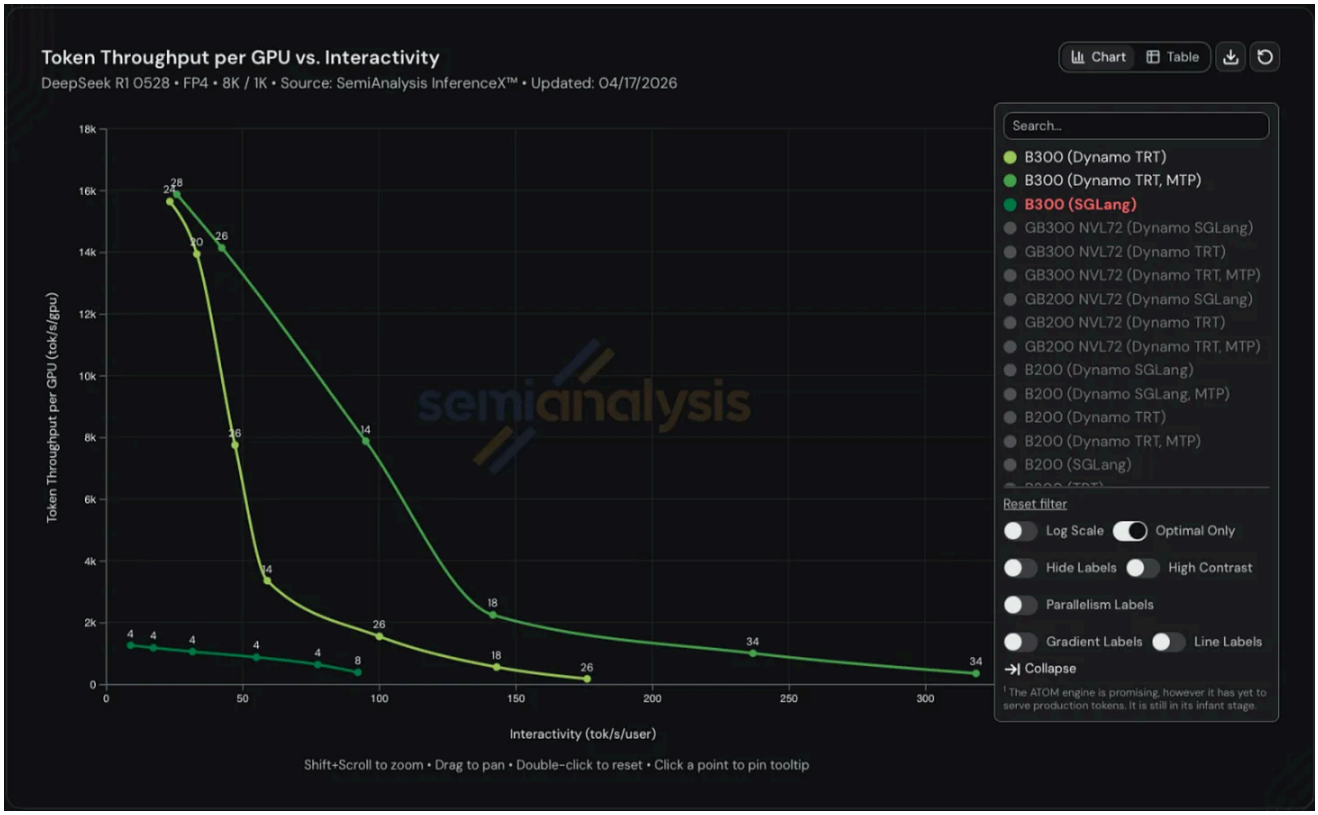
代币（token）的生产成本已大幅下降，因为加速器价格的代际上涨已被更高的吞吐量（tokens/sec/gpu）所抵消。在过去几个月中，平均混合每百万代币价格急剧下降，代理型（agentic）工作负载本质上是多轮对话，具有更长的输入/输出比和更高的缓存命中率，但推理毛利率在同一时期内已从不足 40% 上升至 70% 以上。有关 OpenAI、Anthropic 等所有主流模型每百万代币真实混合价格、代币产量和毛利率的深入估算，请参阅我们的代币经济学（Tokenomics）模型。

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

鉴于硬件和软件的不断改进，InferenceX 仍然是追踪开源模型随时间变化的真实推理性能的最佳基准。

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.

下图显示了在 B300 上运行 DeepSeek R1（输入 8k token，生成 1k token）时，吞吐量与交互性的对比关系。顶部的线条反映了采用 wideEP + disagg + MTP 后的 token 吞吐量，中间线条反映了采用 wideEP + disagg 的情况，而最底部的线条则是在没有这三项软件优化下的表现。这一差距令人震惊：在相同的硬件条件下，同一块 B300 分别能产生约 1k、8k 和 14k tokens/sec/gpu 的性能。仅通过软件改进，就能将吞吐量提升 14 倍。

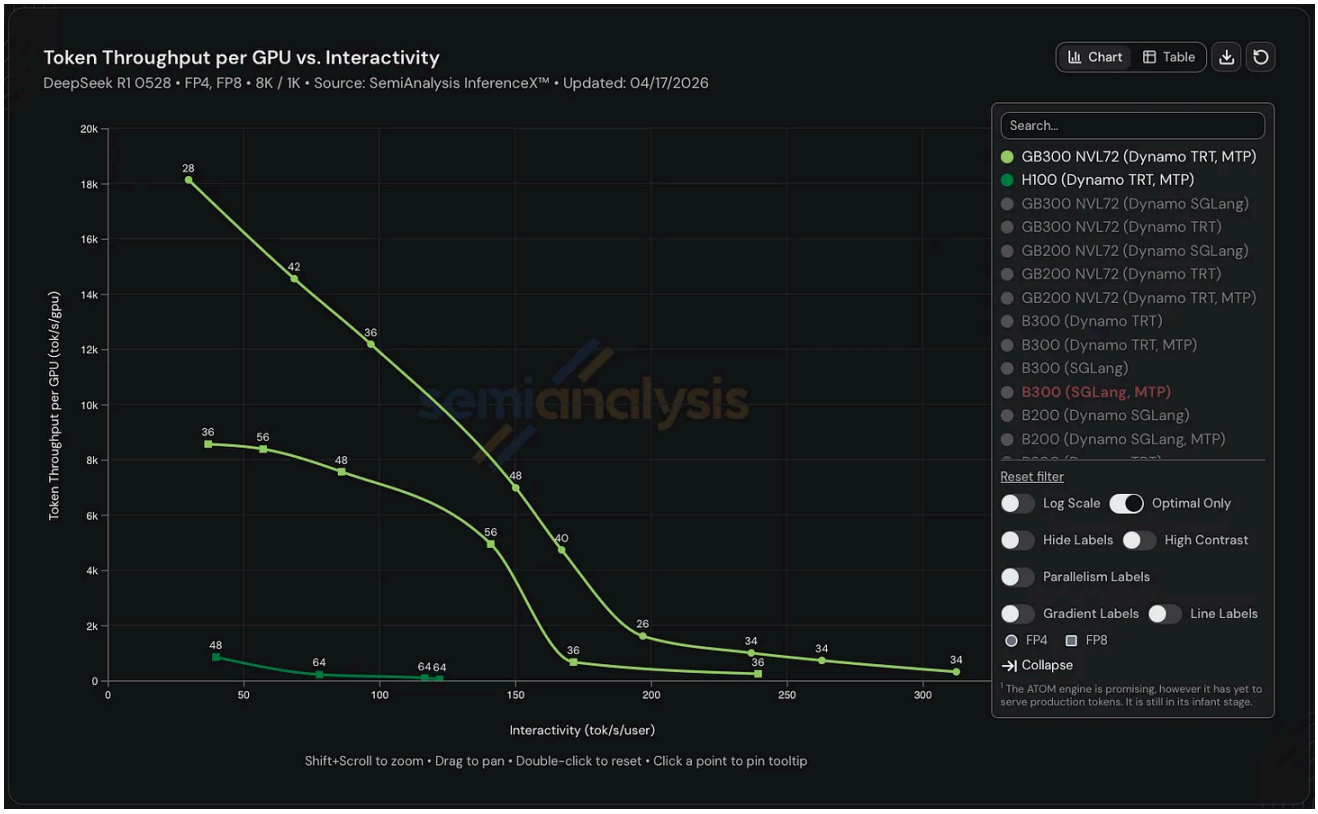


Source: SemiAnalysis InferenceX

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.

如果将硬件改进也考虑在内，这种差异会更加显著。在 FP8 精度下，最优化配置的 GB300 NVL72 的吞吐量比最优化配置的 H100 高出约 17 倍。如果我们切换到 Hopper 架构原生不支持的 FP4 精度，这一差距将跃升至 32 倍。请记住，GB300 的单 GPU 总拥有成本仅比 H100 高出约 70%。



Source: SemiAnalysis InferenceX

SemiAnalysis InferenceX

Model Provider Margins Will Continue to Increase

模型提供商的利润率将持续增长

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.

当 Anthropic 在 2025 年 11 月下旬以每百万输入 Token 5 美元、每百万输出 Token 25 美元的价格发布 Opus 4.5 时，许多人都感到惊讶。此前的 Opus 模型，如 4 和 4.1（分别于 2025 年 5 月和 2025 年 8 月发布），其定价曾高出 3 倍，分别为 15 美元和 75 美元。

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.

然而，我们认为尽管平均售价（ASP）有所下降，但得益于 Trainium 和 Nvidia GPU 的软件优化，以及用 Blackwell 替换了 Hopper，Anthropic 在 Opus Token 上的利润率实际上有所提高。

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.

Anthropic 迄今为止的利润率增长源于成本降低；他们能以更低的成本生成同样的 token。尽管 Opus 进行了降价，但由于大部分业务量从 Sonnet 转向了 Opus，他们的平均销售价格（ASP/token）实际上反而上升了。

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.

即使 XPU 供应商开始大幅提价以更好地获取其在吞吐量提升中的收益份额，Anthropic 仍有另一个杠杆可以用来进一步扩大利润率：他们可以继续将业务量转向更昂贵的 SKU。

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.

正如前文所述，前沿级 Token 的价格与其所能产生的劳动经济价值之间的差距正处于历史最高点。Anthropic 既可以提高 Opus 系列基础款的价格，也可以推出新产品。我们已经看到了后者的出现：Opus fast 的定价是普通 Opus 的 6 倍，而新发布的 Mythos 定价为 $25/$125（是普通 Opus 定价的 5 倍）。这两个 SKU 的利润率都高于普通 Opus，但那些对 AI 接受度最高的企业依然非常乐意支付更高的价格，因为生产力的提升远超其成本。如果 Anthropic 允许我们以 $150/$750 的价格购买 Mythos fast，我们也愿意买单。

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.

前沿模型提供商低毛利率的时代已经结束。真正的智能体 AI（Agentic AI）永久性地提高了每个 Token 的市场出清价格，而且这一趋势不可逆转。

Why Model Provider Profits Won't Get Competed Away

为什么模型提供商的利润不会因竞争而流失

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.

尽管每代币的效用有所增加，但实验室仍无法获得更高利润率，对此最显而易见的论据是竞争。然而，出于两个原因，我们认为情况并不会这样发展。

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.

首先，事实已经证明前沿模型依然保持着定价权。无论基准测试结果如何，在处理实际的知识性工作时，开源模型与闭源模型相比仍有明显差距，且没有理由认为这种差距会在短期内缩小。Kimi K2.6（$0.95/$4）对 Opus 的定价几乎没有产生下行压力。

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.**

其次，算力限制意味着没有任何一家顶尖实验室能够服务整个市场。Anthropic 如今已经开始疏远大批市场用户，其做法是将 Claude Code 锁定在每月 100 美元以上的订阅门槛后，并封杀 OpenClaw 等第三方工具。在可预见的未来，Token 需求将远超供给，这意味着任何能够提供真正顶尖质量的实验室，都将能够根据 Token 交付的经济价值来定价，而不是通过竞争耗尽彼此的利润空间。

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

智能体 AI（Agentic AI）冲击市场，但台积电和英伟达依然稳如泰山

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.

尽管黄仁勋在最近的 GTC 主旨演讲中反复强调智能体 AI（agentic AI），但 Nvidia 和台积电仍未完全领悟过去几个月代币经济学（token economics）所发生的变革性影响。从黄仁勋对 InferenceX 的反应来看，我们已经发现 Nvidia 低估了 Blackwell 在性价比方面的提升，而现在看来，他们似乎也低估了前沿代币价值升值的速度。

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.

Nvidia 仍在一个由旧假设塑造的框架内运行，即单位算力的付费意愿会随时间推移而下降。但这一假设已不再成立。受代理型工作负载（agentic workloads）爆发和单个工作流 Token 消耗量剧增的驱动，市场已发生实质性转变。需求不再是线性增长，而是呈复利式增长。

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.

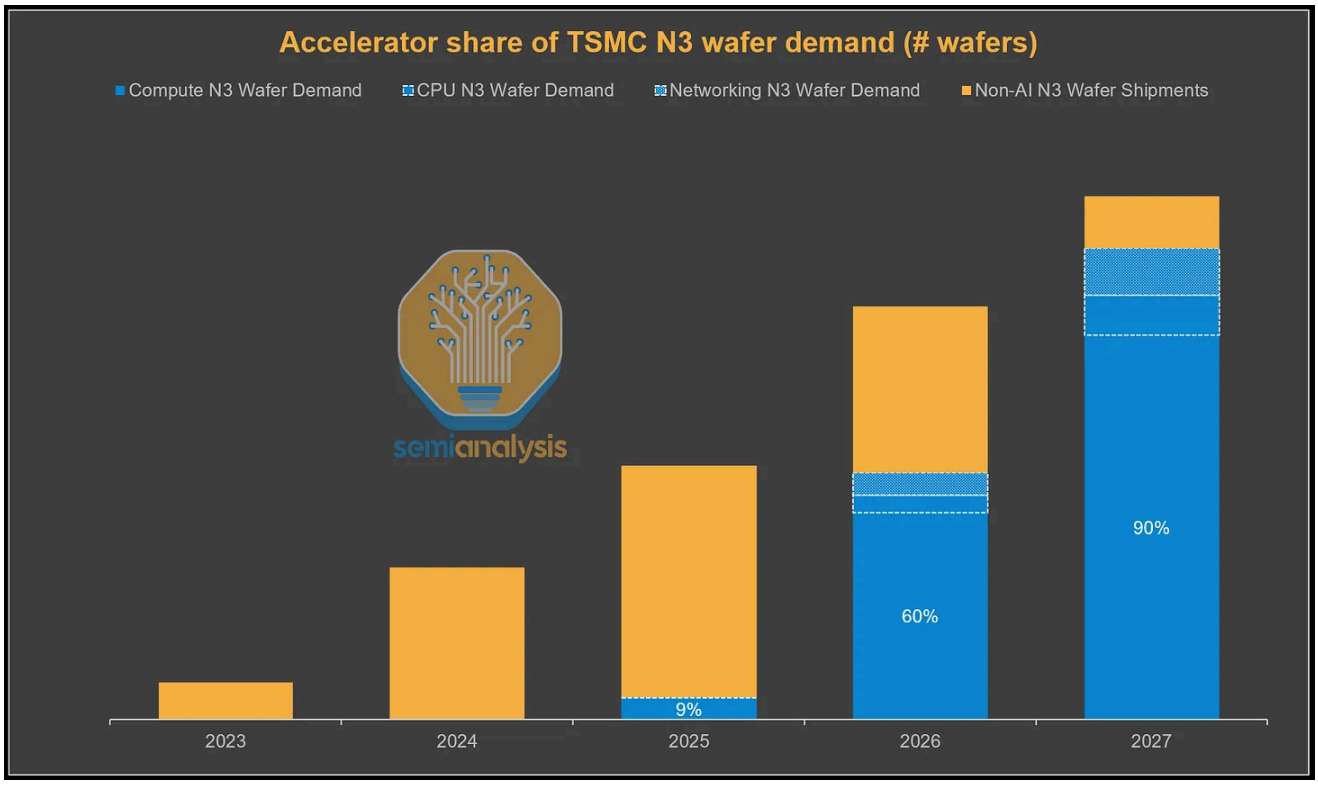
然而，需求仍在持续加速。据报道，Anthropic 的年度经常性收入（ARR）已达到 440 亿美元以上，高于我们上次更新时的 300 亿美元；同时，GLM 和 Kimi 等权重开放模型正在扩大可触达的算力基数。AI 实验室和新型云服务商（neoclouds）的融资正直接转化为新增的 GPU 部署。

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.

与此同时，算力供应在结构上依然受限。内存和先进制程晶圆的上游瓶颈持续限制着供应能力，预计 N3 产能利用率将在 2026 年下半年超过 100%，而 DRAM 晶圆厂的利用率已超过 90%。目前看不到任何实质性的缓解迹象。

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.

台积电本可以大幅提高价格，但他们并没有这样做。这是他们犯下的一个战略性错误。如果不涨价，他们至少可以要求更多的预付款。



Source: SemiAnalysis Foundry Model, SemiAnalysis Accelerator Model

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.

当前计算市场的动态表明，如果趋势延续，由巨大的代币终端需求所产生的价值将继续流向 AI 实验室、超大规模云服务商、推理提供商、新型云服务商（Neoclouds）以及内存供应商。

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.

在强劲的终端需求、不断增长的 Token 变现能力以及日益优化的单位经济效益推动下，AI 实验室正获取不成比例的价值份额。与此同时，尽管 Nvidia 的硬件仍是实现这一价值创造的关键瓶颈，但其定价框架尚未完全调整以反映这一转变。尽管 Token 变现能力在提升且单位经济效益日益向好，Nvidia 算力依然是实现这些价值创造的基石。

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.

各层级客户对 Nvidia 系统的需求依然极其强劲，买家愿意锁定长期合同并接受更高定价以确保产能。即便存在其他硬件选项，Nvidia 在生态系统成熟度、软件栈和部署可靠性方面仍保持着明显优势。对于许多工作负载，尤其是前沿领域，替代品尚无法完全互换。

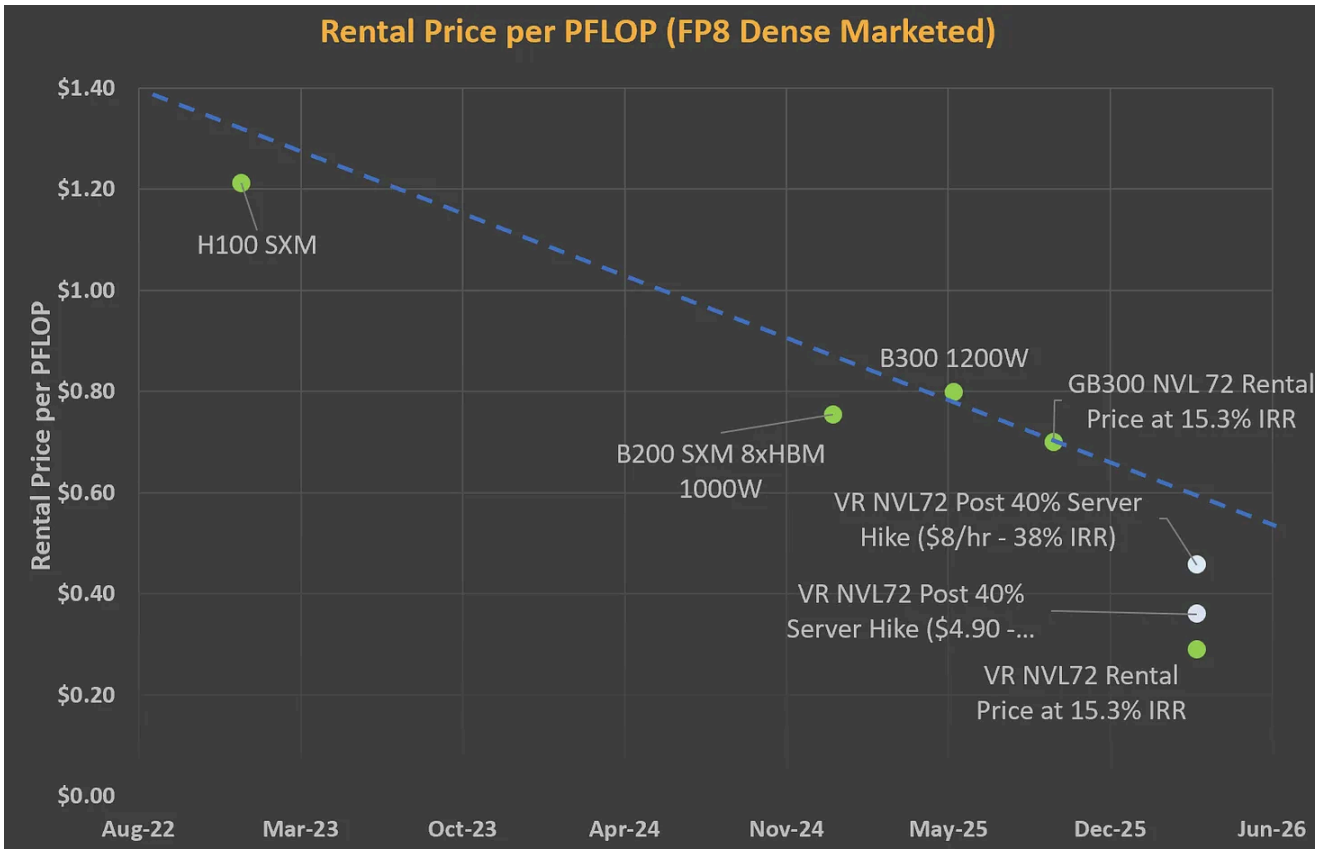
Rubin, set to launch in 2H26, sits at the center of these dynamics. It delivers a step-function improvement in performance but also embeds a much larger memory subsystem at a time when memory is the tightest constraint in the supply chain. DRAM pricing has already moved sharply higher and is likely to remain elevated, making memory the primary driver of system cost.

Rubin 计划于 2026 年下半年发布，处于这些动态的核心位置。它不仅实现了性能的阶梯式提升，还在内存成为供应链最紧绷约束的时刻，嵌入了规模大得多的内存子系统。DRAM 价格已经大幅上涨，并可能持续保持高位，这使得内存成为系统成本的主要驱动因素。

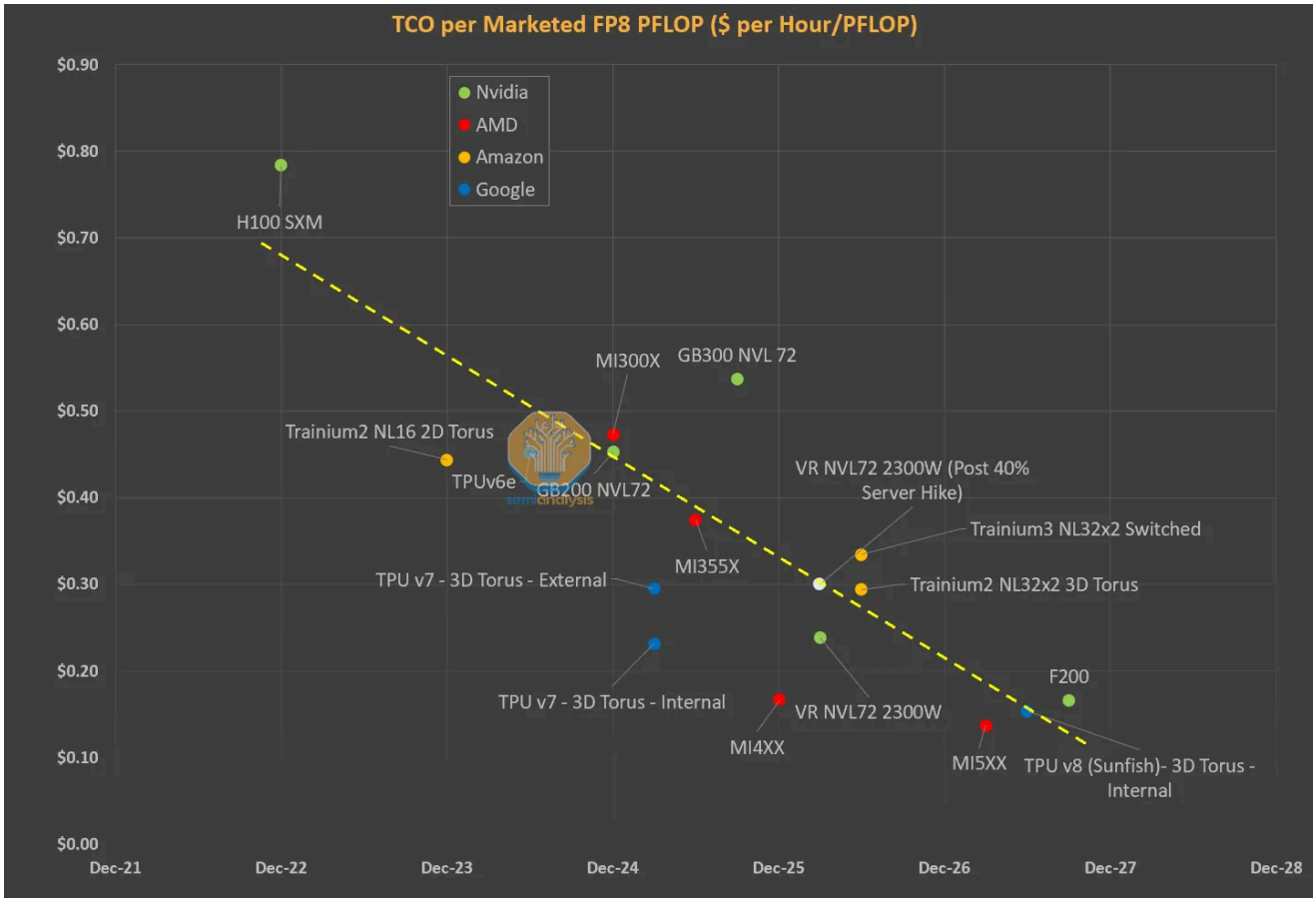
In this context, Nvidia has room to increase pricing, particularly for systems like Rubin that deliver step-function performance gains. The incremental value created at the system level far exceeds the incremental cost, especially when viewed through

$/FLOP or end workload economics.

在这种背景下，Nvidia 仍有提价空间，特别是对于像 Rubin 这样能带来阶跃式性能提升的系统。在系统层面创造的增量价值远超其增量成本，尤其是从单位算力成本（$/FLOP）或终端工作负载经济效益的角度来看。



Source: AI TCO Model AI TCO 模型



Source: AI TCO Model 来源：AI TCO 模型

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.

这造成了明显的脱节。市场已经发生了结构性转变，需求的增长速度和持续性远超供应的响应能力。然而，Nvidia 的定价框架仍锚定在先前的假设上，未能根据其系统目前所交付的更高价值进行相应调整。

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.

简而言之，即使 Nvidia 提高服务器价格，基础设施提供商提高算力价格，需求仍将保持不变。买家正在优化算力获取渠道，终端用户正在优化尽可能多的 Token 获取量，双方都在不惜一切代价确保产能——边际成本优化并非他们当下的首要考量。

SOCAMM Pricing: Nvidia's Next Margin Lever

SOCAMM 定价：Nvidia 的下一个利润杠杆

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

继 Nvidia 能否涨价之后的下一个问题是，在系统内部的哪个环节涨价最为有效。

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.

在系统层面，内存是最自然的控制点。Rubin 级系统在已经受限的供应链中嵌入了显著更多的内存，且与算力不同，内存可以更清晰地进行细分并持续重新定价。

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.

这是因为 GB200 NVL72 上的内存是一种基于插槽式 LPDDR 的内存解决方案，称为 SOCAMM（片上系统附加内存模块）。SOCAMM 专为 Nvidia 的机架级系统设计，能够实现更高的容量、模块化、能效，并允许内存与计算资源独立定价。

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.

这使得 SOCAMM 成为理解 Nvidia 定价策略最重要的变量之一。两个因素最终决定了系统级的定价结果：Nvidia 采购 SOCAMM 的成本，以及将该内存转售给客户时施加的加价。鉴于其“极致共同设计”方法的复杂性以及错综复杂的供应链动态，要对 Nvidia 机架级系统的定价和物料清单（BoM）建立精确的视图并非易事。

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:

这就是为什么 SemiAnalysis 通过我们的 VR NVL72 物料清单（BoM）和功耗预算模型提供了行业领先的详细分析。此外，在确定面向终端客户的内存定价时，有两个关键的波动因素在起作用：

1. The price Nvidia secured for SOCAMM2, and

   Nvidia 为 SOCAMM2 锁定的价格，以及

2. The markup Nvidia applies to SOCAMM when selling to customers,

   Nvidia 在向客户销售时对 SOCAMM 施加的加价，

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

两者都是影响最终给客户报价的关键因素。

	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

SemiAnalysis 内存模型

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

截至今日，我们的内存模型显示 Nvidia 支付的 SOCAMM 合约价格在 2026 年第一季度约为 8 美元/GB，较 2025 年第四季度大幅上涨。这一跳涨是由第一季度 LPDDR5X 价格的普遍飙升以及内存供应的整体收紧所驱动的。我们基于以下两点得出这一预估：

1. SOCAMM should price at a premium to mobile LPDDR5X (~$6–7/GB in 1Q26) given higher development complexity and longer cycle times.

   考虑到更高的开发复杂性和更长的周期时间，SOCAMM 的定价应高于移动端 LPDDR5X（2026 年第一季度约为 6-7 美元/GB）。

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.

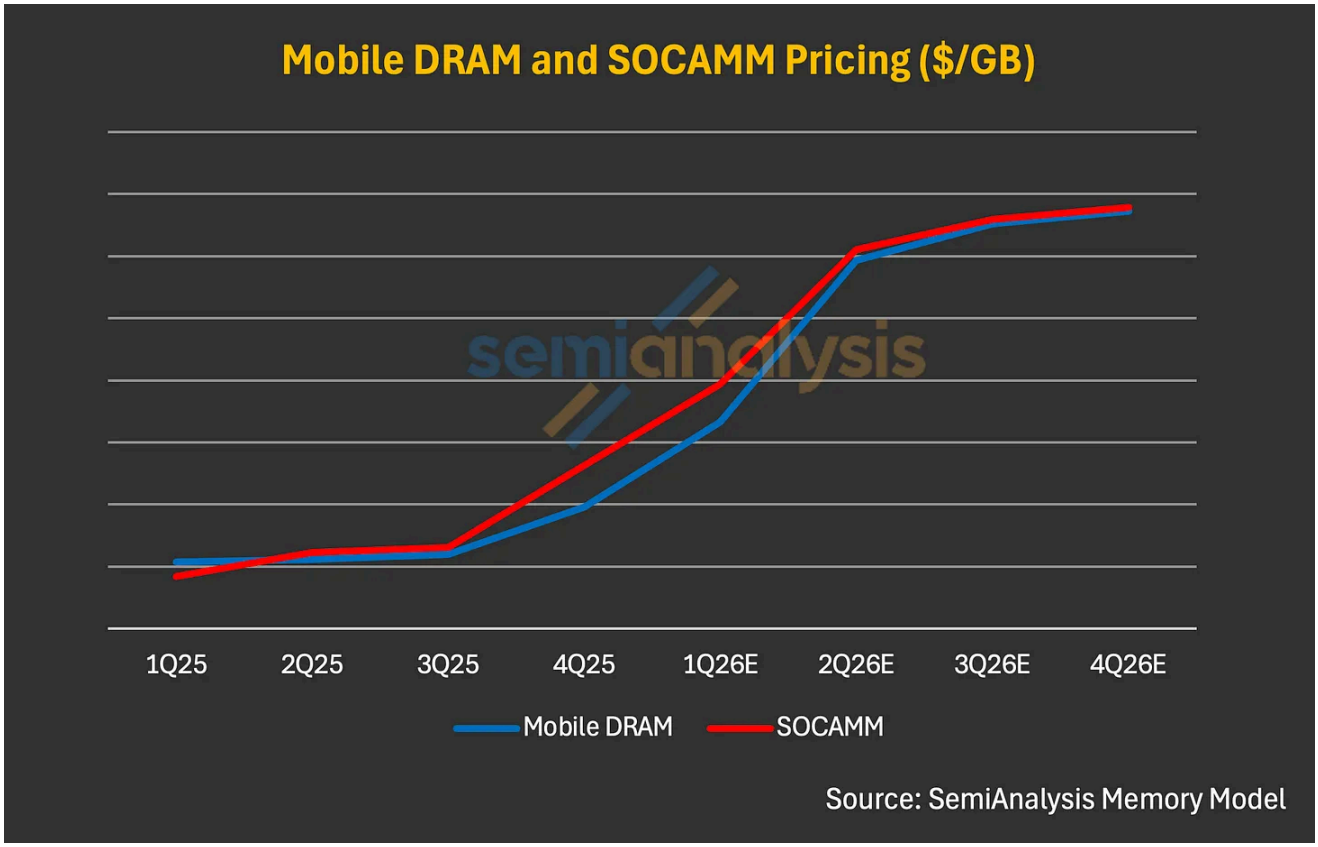
   移动端 LPDDR5X 价格的上涨应会在同一时期传导至 SOCAMM，因为受限的 LPDDR5X 以及更广泛的大宗商品 DRAM 供应是由消费级和服务器需求共同瓜分的。

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.

行业传闻表明，Nvidia 已为其 GB300 NVL 72 和 VR NVL72 系统锁定了大量的 SOCAMM 产能，采用的是长期协议（LTA）模式，这一点我们在早前针对内存模型的机构报告中已有概述。作为目前唯一的规模化 SOCAMM 客户，且可以域最关键的买家，Nvidia 可能会从优先供应和定价中获益；我们认为，在能力方面，Nvidia 过去的往绩记录已足以说明一切。

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.

即便如此，更广泛的 DRAM 价格动态仍将不可避免地产生影响。未来几个季度移动端 LPDDR5X 价格的进一步上涨仍将是 SOCAMM 的关键定价参考，鉴于 LPDDR5 分配量有限，SOCAMM 应据此重新定价。我们认为 SOCAMM 在 26 年底的定价可能会超过 13 美元/GB，这与预计今年年底的移动端 DRAM 定价大致持平；因此，我们认为约 10 美元/GB 是 Nvidia SOCAMM 成本的一个合理假设。



Source: SemiAnalysis Memory Model

来源：SemiAnalysis 内存模型

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:

一些人可能会提出的一个关键问题是：客户凭什么接受 Nvidia 进一步的涨价和利润率扩张，而 Nvidia 又能以什么理由令人信服地证明这种立场的合理性？我们认为 Nvidia 在 SOCAMM 上收取 60% 的利润率是合理的，原因有三：

- 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.

  首先，当前的环境对 Nvidia 非常有利。各处的内存供应都处于受限状态，而 Nvidia 与其客户及同行竞争对手相比，已经锁定了最大的供应量（至少在 SOCAMM 方面），这将使该公司能够利用这一供应链优势。

- 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.

  其次，就性能功耗比（Performance per TCO）而言，VR NVL72 仍是目前市场上表现最好的平台，且该系统的生产拥有复杂但成熟的供应链支撑。为了实现计算投资收益的最大化，客户可能别无选择，只能接受英伟达的新定价模式。

- 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.

  最后，由于英伟达作为 SOCAMM2 的采购方，本身就面临着实质性的价格上涨，我们认为假设客户会接受英伟达在 SOCAMM2 成本之上收取的毛利，对于 VR NVL72 来说是合理的。

Capex Per Watt Trends from GB300 to VR NVL72

从 GB300 到 VR NVL72 的每瓦资本支出趋势

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.

对于 GB300，DRAM 被捆绑到板卡中，并以约 75% 的毛利率加价，这使得板载内存收取的利润率与 Blackwell 系统隐含的定价保持一致。

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.

对于 Rubin，我们最初假设了同样的动态，即 Nvidia 将目标定在 70% 左右的整体系统毛利率。因此，我们最初的物料清单 (BoM) 模型在整个 Strata 板卡中应用了一致的利润率，使 SOCAMM 的利润率也保持在 70% 左右。

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.

然而，由于 SOCAMM2 在 Rubin 中是一个插槽式模块，而 GB300 使用的是焊接在板上的普通 LPDDR5X 模块，因此内存可以从基础系统中解耦并单独报价。这使得 Nvidia 能够将内存明确地作为独立项目进行定价，而不是将其嵌入到板级定价中。更重要的是，这还为 Nvidia 提供了一个额外的价值维度，使其能够在保持主板利润率不变的同时，灵活调整 SOCAMM2 的利润率。即使 Nvidia 最初承担了部分内存成本上涨的压力，它仍有能力通过向客户收取更高的系统级利润来抵消这一影响。

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.

因此，我们原以为随着从 GB300 过渡到 VR NVL72，每瓦特的总资本支出会上升。然而事实恰恰相反——目前的定价换算下来，每瓦特资本支出仅从 GB300 的 37.4 美元/瓦轻微攀升至 VR NVL72 的 38.1 美元/瓦。尽管从 GB300 到 VR NVL72，芯片的 TDP 几乎翻了一番（从 1400W 增加到 2300W），且算力（FLOPs）也有实质性的提升，但情况依然如此。

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

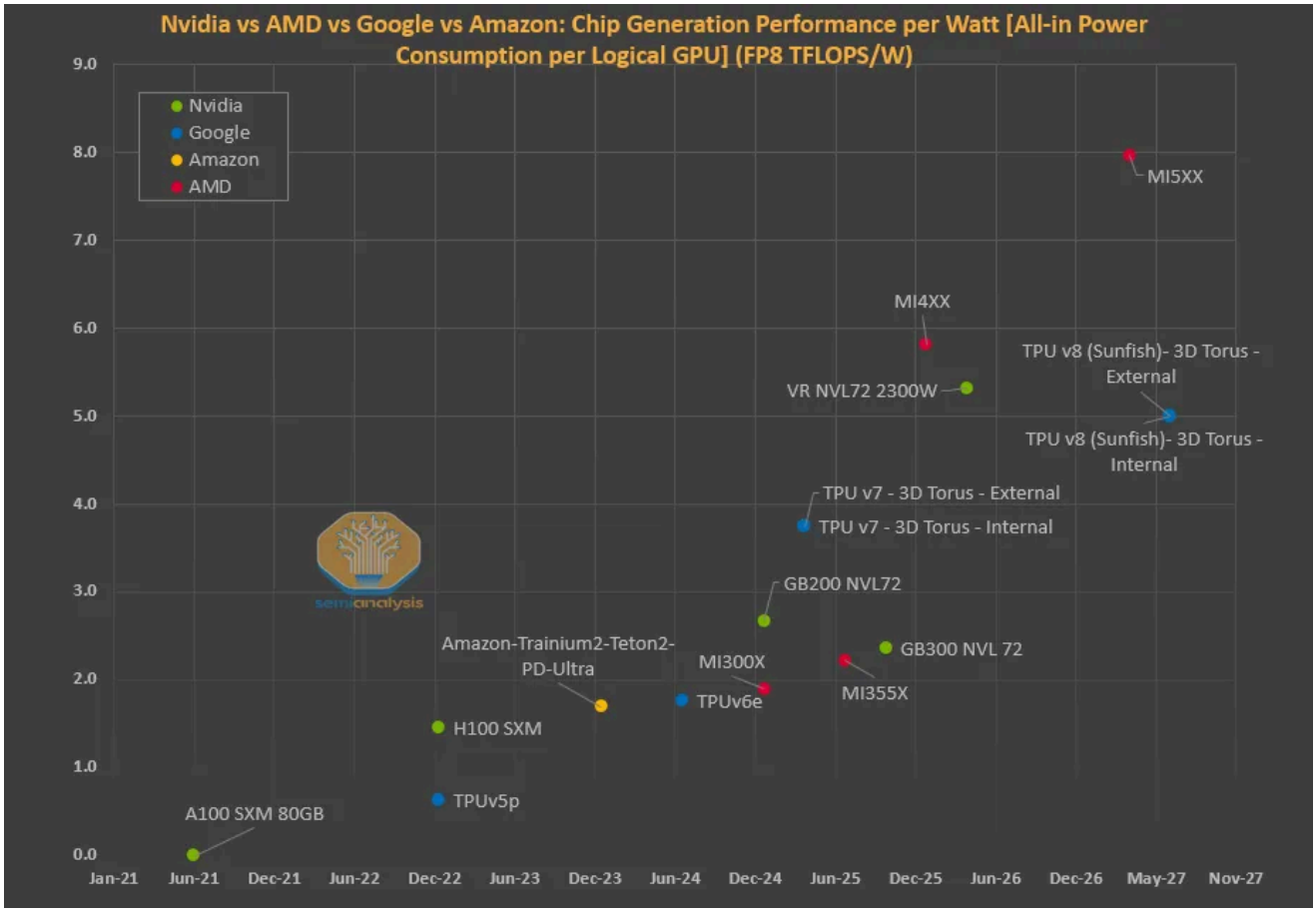
SemiAnalysis AI TCO 模型

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.

相对于每瓦服务器资本支出的更广泛趋势而言，这是极不寻常的。在 AMD、Nvidia 和定制 ASIC 领域，每瓦资本支出通常会随着代际更迭而增加，因为每瓦性能的提升允许供应商在系统层面获取更多价值。因此，令我们感到困惑的是，从 GB300 到 VR NVL72，每吉瓦成本（$/GW）似乎基本保持停滞。考虑到从 GB300 到 VR NVL72 的每瓦性能提升了一倍以上，这种情况就显得更加反常了。

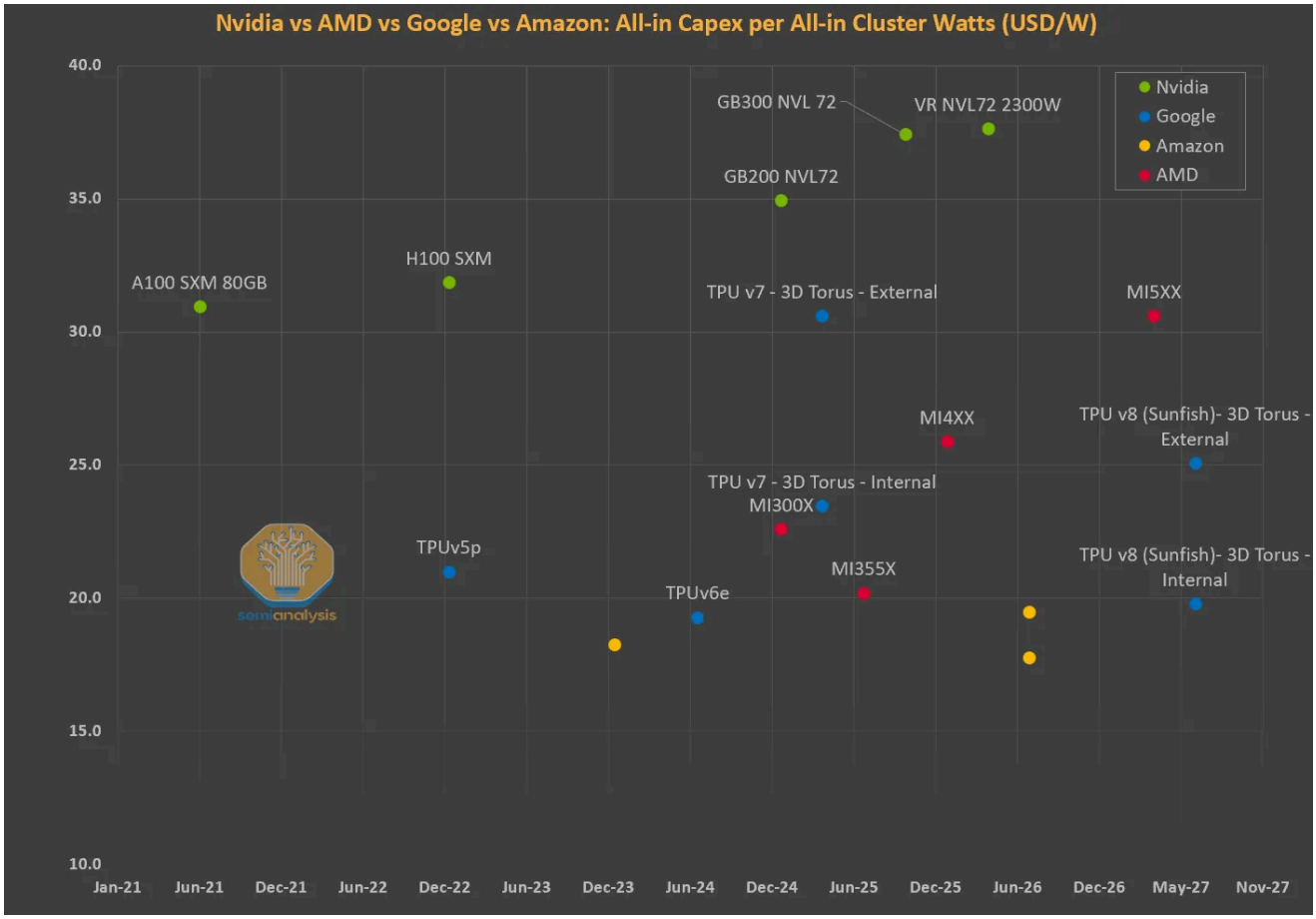
Nvidia also has the opporunity to price discriminate on memory more than they do on the GPU because memory isn’t an anti-trust concern whereas the GPU is.

Nvidia 还有机会在显存上比在 GPU 上进行更大程度的价格歧视，因为显存不涉及反垄断问题，而 GPU 则涉及。



Source: SemiAnalysis AI TCO Model

SemiAnalysis AI TCO 模型



Source: SemiAnalysis AI TCO Model

SemiAnalysis AI TCO 模型

Networking as a Vector for Price Discrimination

网络作为价格歧视的媒介

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.

如今，Nvidia 在同一层级的不同客户之间并不会对 GPU 价格进行大幅区分。核心组件在整个生态系统中的售价通常基本一致，无论是面向超大规模云服务商（Hyperscalers）、新型云服务商（Neoclouds）、新兴型云服务商、主权国家还是企业。这为核心 GPU 和内存组件创造了一个相对统一的价格结构，即便是在支付意愿差异巨大的市场中也是如此。

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.

虽然 GPU 的定价相对统一，但 Nvidia 传统上在网络设备方面采取价格歧视策略，向 Neoclouds 和其他边缘云厂商提供的价格点远高于超大规模云服务商（hyperscalers）。我们对 GPU 云供应商进行了独立调查，发现例如 SN5610 对 Neocloud 的定价可能是超大规模云服务商的两倍。超大规模云服务商显然拥有更强的议价能力，但这并不是因为他们从 Nvidia 购买了更多的交换机和收发器。

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.

Neocloud 缺乏定制化和成本优化其网络集群所需的规模和网络专业知识，因此他们最终更倾向于 Nvidia 的开箱即用解决方案。超大规模运营商（Hyperscalers）则直接与 OEM 和 ODM 合作，并拥有深厚的网络工程储备，能够部署更具成本效益的解决方案，而这些方案可能并非开箱即用的部署，因此妥善部署的难度也更大。

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.

然而，从集群总资本成本的角度来看，Neocloud 与超大规模云服务商（hyperscaler）之间网络成本的差异变得不再那么显著。对于两个同类集群，Neocloud 的网络成本比超大规模云服务商高出 94%，但反映在整机柜级服务器的全额资本成本上，仅相当于增加了 10%。这还不包括电力、公用事业和运营等其他变量，这些变量将进一步削弱由网络设备价格差异所导致的成本差距。

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		84%		76%	0%
Optical Transceivers		4%		8%	140%
Switches		6%		9%	64%
Fiber, Cables, Others		1%		2%	96%
Networking Cost		11%		19%	94%
All Others		5%		5%	0%
All-in Cost per Rack-Scale Server		**100%**		**100%**	**10%**
Costs for a Spectrum-X cluster using Nvidia SN5610 Back-end switches with 64 ports of 800G					

Source: SemiAnalysis AI Networking Model

SemiAnalysis AI 网络模型

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.

尽管这是一个展示 Nvidia 如何根据价值为其解决方案定价的绝佳案例研究，但目前 Neoclouds 与超大规模云厂商（Hyperscalers）之间的价格差距已经非常显著，这使得 Nvidia 进一步动用这一杠杆的空间变得有限。

Nvidia as the Central Bank of AI

Nvidia：AI 领域的中央银行

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

英伟达至今在定价方面保持克制的一个原因，可能是监管层面的顾虑与战略上的谨慎共同作用的结果。

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.

鉴于 Nvidia 在 GPU、互连和软件领域的统治地位，其在 AI 计算堆栈中的地位已经面临日益严格的反垄断审查。在这种环境下，如果为了完全捕获所交付的价值而激进地对系统重新定价，可能会招致进一步的关注，特别是如果这导致在下游 AI 实验室也产生巨额利润的同时，自身利润率出现过度扩张。将定价维持在接近先前框架的水平，有助于避免在供应受限的市场中释放出定价权过大的信号。

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.

这种行为并非没有先例。台积电（TSMC）在历史上也采取过类似的做法。即便在满负荷运转并成为先进制程供应瓶颈的情况下，台积电通常也会避免完全根据稀缺性来定价。相反，它优先考虑长期关系和生态系统的稳定性，而非榨取最大的短期利润，这在一定程度上是为了避免监管压力和客户的反弹。

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.

Nvidia 似乎也在遵循类似的路径。它并没有为了反映性能提升和内存成本的结构性变化而对 Rubin 系统进行全面的重新定价，而是保持了一种更为审慎的定价策略。这种做法在利润率扩张与监管风险、生态系统动态以及避免加速客户向替代计算平台转移的需求之间取得了平衡。

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.

我们在《英伟达作为中央银行》一文中也表达了类似的观点。英伟达正积极支持更广泛生态系统的发展，确保长期的需求扩张，而非追求短期利益榨取的最大化。如今，前沿实验室受益于英伟达软件驱动的效率提升，但这些改进并未在硬件层面完全变现。因此，尽管英伟达是主要的推动者，增值收益仍持续流向下游。通过"抽走房间里的氧气"，英伟达旨在确保自己在可预见的未来始终是 AI 时代的主角。

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?

然而，在算力需求远超供应的情况下，那些掌控稀缺资源的人为何不通过提高定价来获取更高的利润呢?

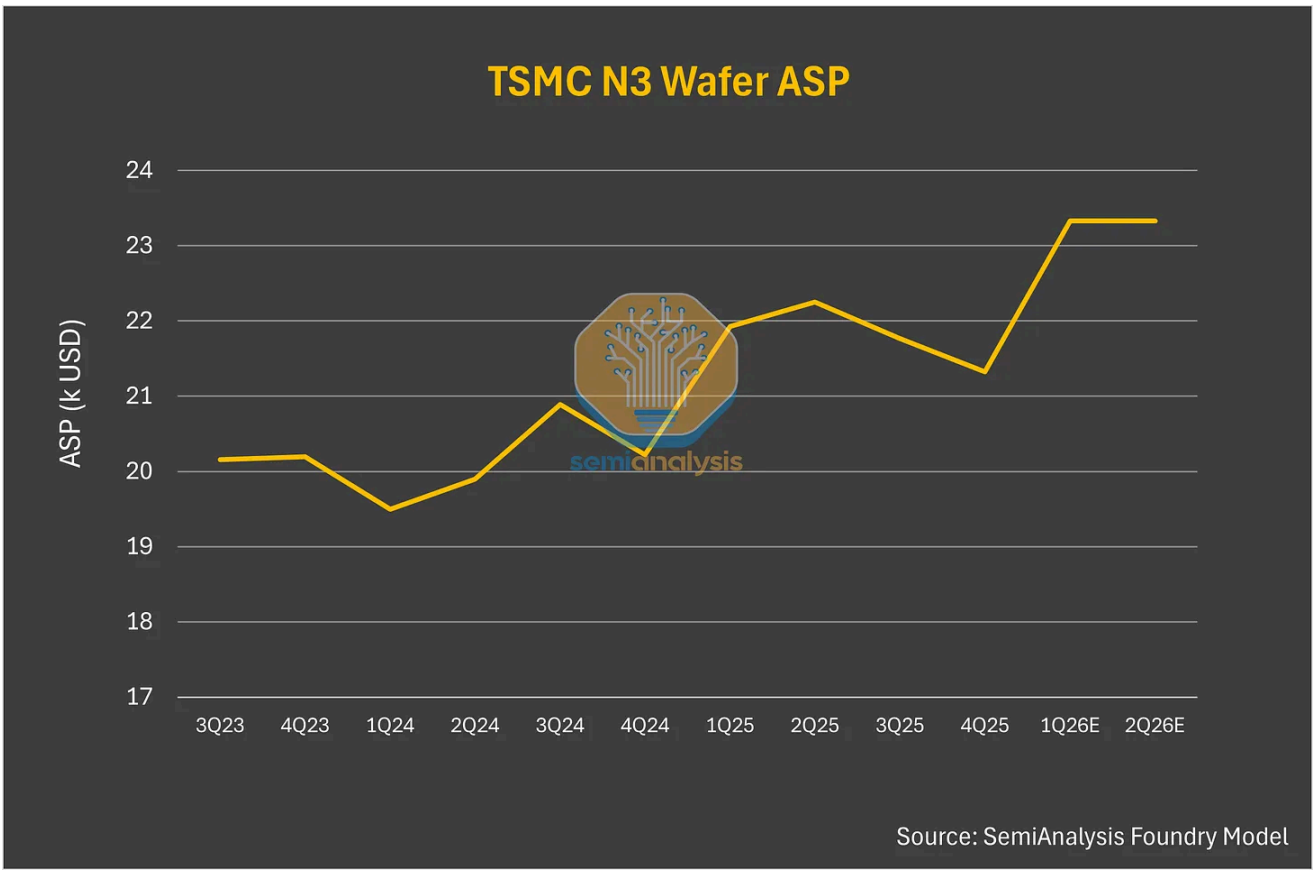
TSMC, The Fairest and Most Just Company In The World

台积电：世界上最公平、最公正的公司

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,

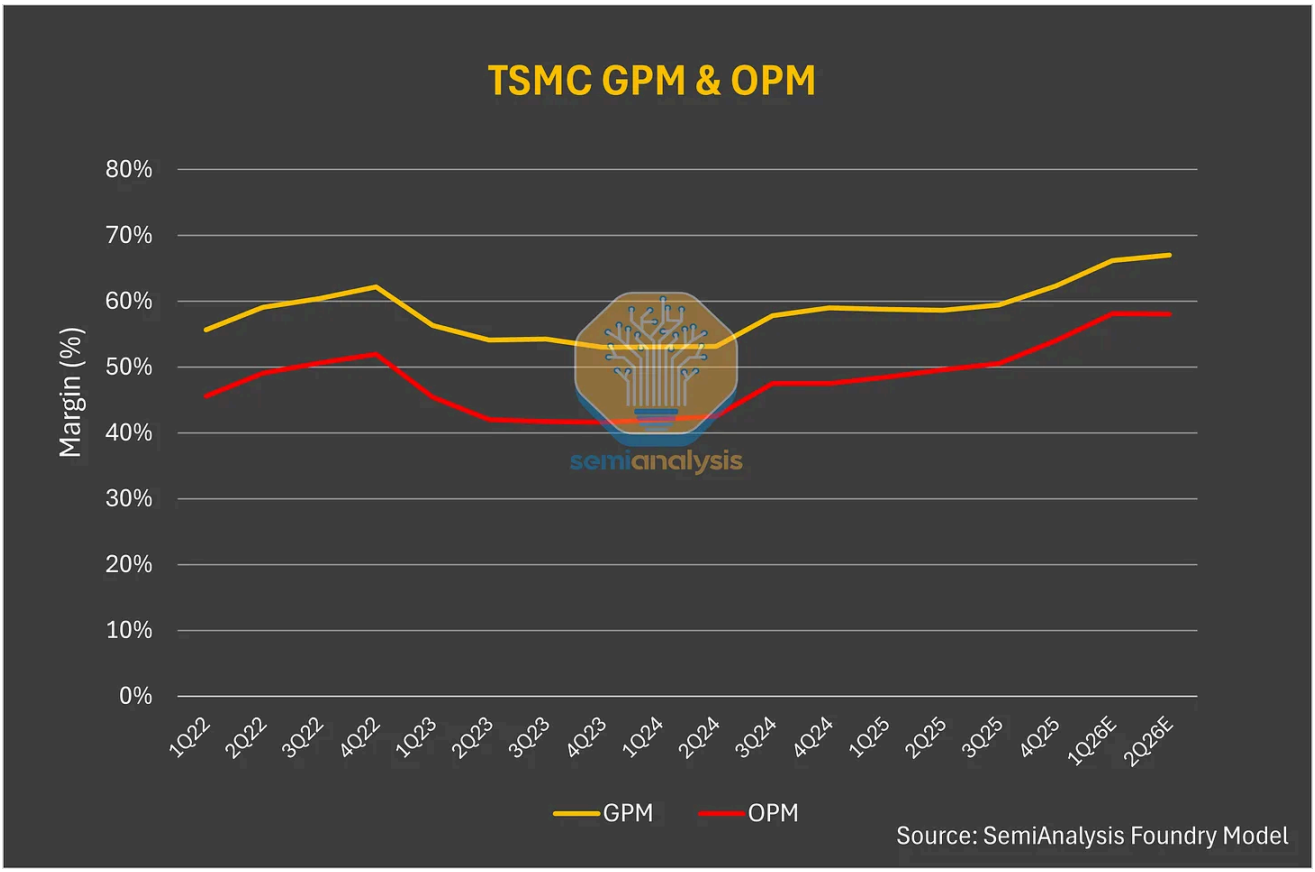
Broadcom, Annapurna, MeidaTek and AMD are all fighting for more N3 wafer allocation from TSMC so that they can ship more compute to their customers. While N3 capacity is arguably the tightest constraint in the system, pricing remains relatively stable.

我们曾提到台积电的 N3 产能甚至更加紧张。目前，所有主要的加速器路线图都已在今年和明年汇聚于 N3 工艺节点。Nvidia、Broadcom、Annapurna、MediaTek 和 AMD 都在争夺台积电更多的 N3 晶圆配额，以便向客户交付更多算力。尽管 N3 产能可以说是系统中最为紧绷的约束条件，但其定价仍保持相对稳定。



Source: SemiAnalysis Foundry Industry Model

来源：SemiAnalysis 晶圆代工行业模型



Source: SemiAnalysis Foundry Industry Model

来源：SemiAnalysis 晶圆代工行业模型

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.

台积电在战略上倾向于在行业下行周期内保护其盈利能力。而这一政策的另一面是，它也削弱了在上行周期中的增长空间。无论如何，台积电无疑未能赚尽所有潜在利润，因为其主要的无晶圆厂客户都享有极高的毛利率，而这些利润本可以转移给台积电。

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.

然而，台积电完全可以采取更激进的定价策略，客户不仅会接受，我们甚至认为有些客户还会表示欢迎。如果支付更高的晶圆价格意味着能将那些支付能力较弱的竞争对手拒之门外，Nvidia 会非常乐意效劳。毕竟，黄仁勋本人在 2024 年就曾表示台积电应该提高晶圆收费，而他这么说正是出于这个原因。

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.

台积电也可以采取长期协议的立场，以产能保证承诺和预付款来代替大幅提价。这是一种更有可能的路径。

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

我们认为 Nvidia 正开始变得越来越像 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.

在这种环境下，它最大的优势在于采购。Nvidia 已经获得了不成比例的受限上游供应访问权，特别是 TSMC 晶圆，这使其能够满足其他公司无法满足的需求。

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.

因此，像 Anthropic 这样的 AI 算力买家被迫进入 Nvidia 的生态系统，因为来自 TPU 和 Trainium 的替代产能仍受限于同样的上游瓶颈。尽管拥有这种结构性优势，Nvidia 并没有在定价中充分体现这一点。

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.

目前，Nvidia 的定价仍锚定在基于成本的框架上。但这不太可能持久。随着推理提供商的投资回报率变得更加清晰并得到更广泛的认可，焦点将进一步转向按价值定价。这将减少对定价的审视，并为 GPU 基础设施提供商从基于成本的定价转向基于价值的定价提供空间。一旦这种转变发生，它将为 Nvidia 提高价格并捕获更多系统级交付的价值创造空间——这也是利益蛋糕不断扩大的体现。

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

推算 VR NVL72 租赁定价：基于成本与基于价值的方法论

There are two main approaches to pricing:

定价主要有两种方法：

1. Cost-based Pricing and, 基于成本的定价，以及
2. Value Based Pricing. 基于价值的定价。

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.

基于成本的定价方法始于这样一个前提：只有当 GPU 部署项目能够满足 Neoclouds 的最低回报阈值时，这些项目才会启动。如果回报率低于这一水平，在价格调整到足以跨越该门槛之前，将不会部署任何产能。

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.

因此，在基于成本的框架下收取的租赁价格，是使 Neocloud 获得的项目内部收益率（IRR）高于部署最低门槛收益率的价格。目前大多数项目的回报率往往处于 15% 到 19% 之间。以目前的 GB300 部署为例，在预付 15% 的情况下，5 年期的项目内部收益率可能达到 15.6%。

Scenario Settings

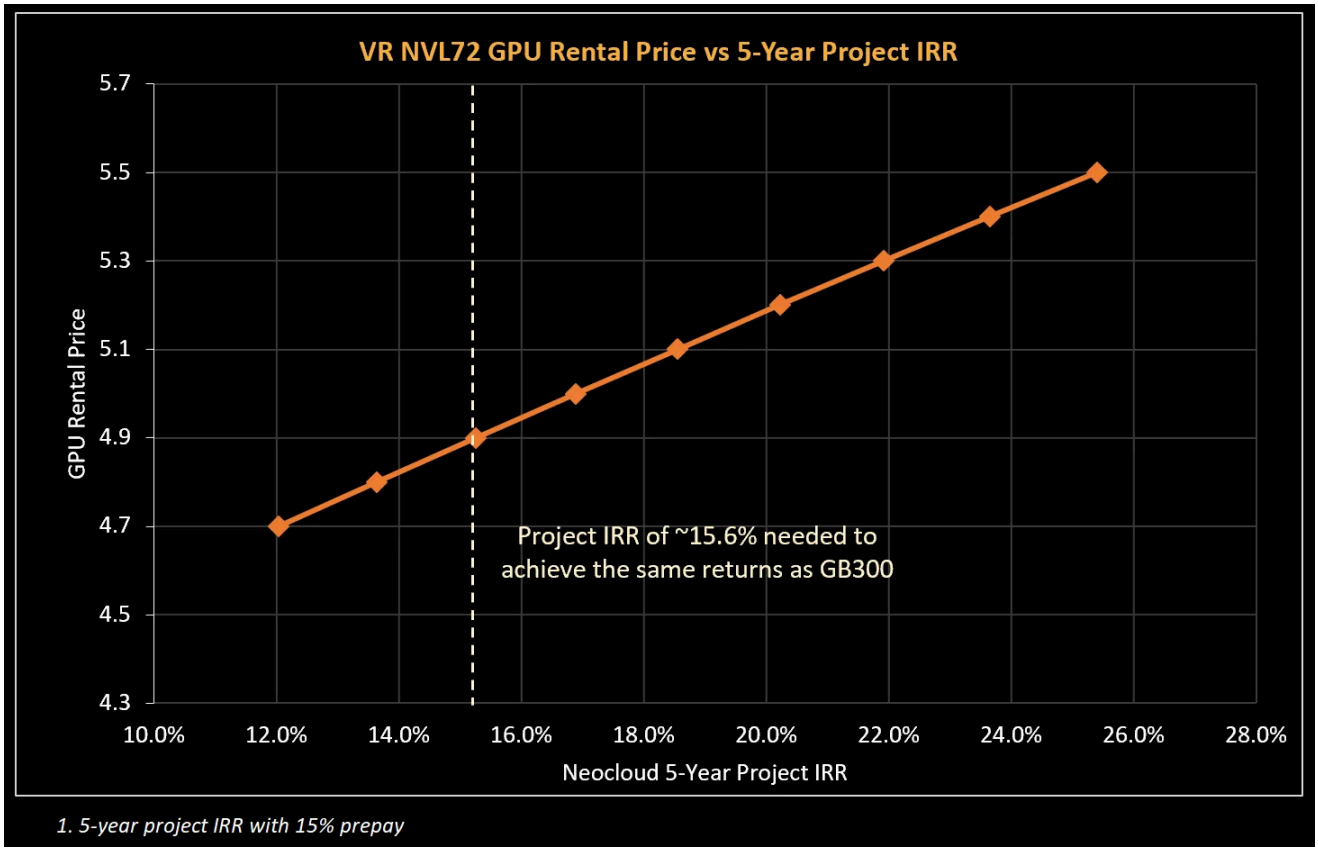
System Configuration	GB300 NVL 72	[redacted]
Number of Accelerators in Project	72	
Price per Server	[redacted]	**USD**
Logical GPUs per Server	72	
Number of Servers	[redacted]	
Total Server Capex		**USD**
System First Production Date	**30-Sep-25**	
Cost of Capital for NPV	[redacted]	
Customer Prepay %		
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

SemiAnalysis AI TCO 模型

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.

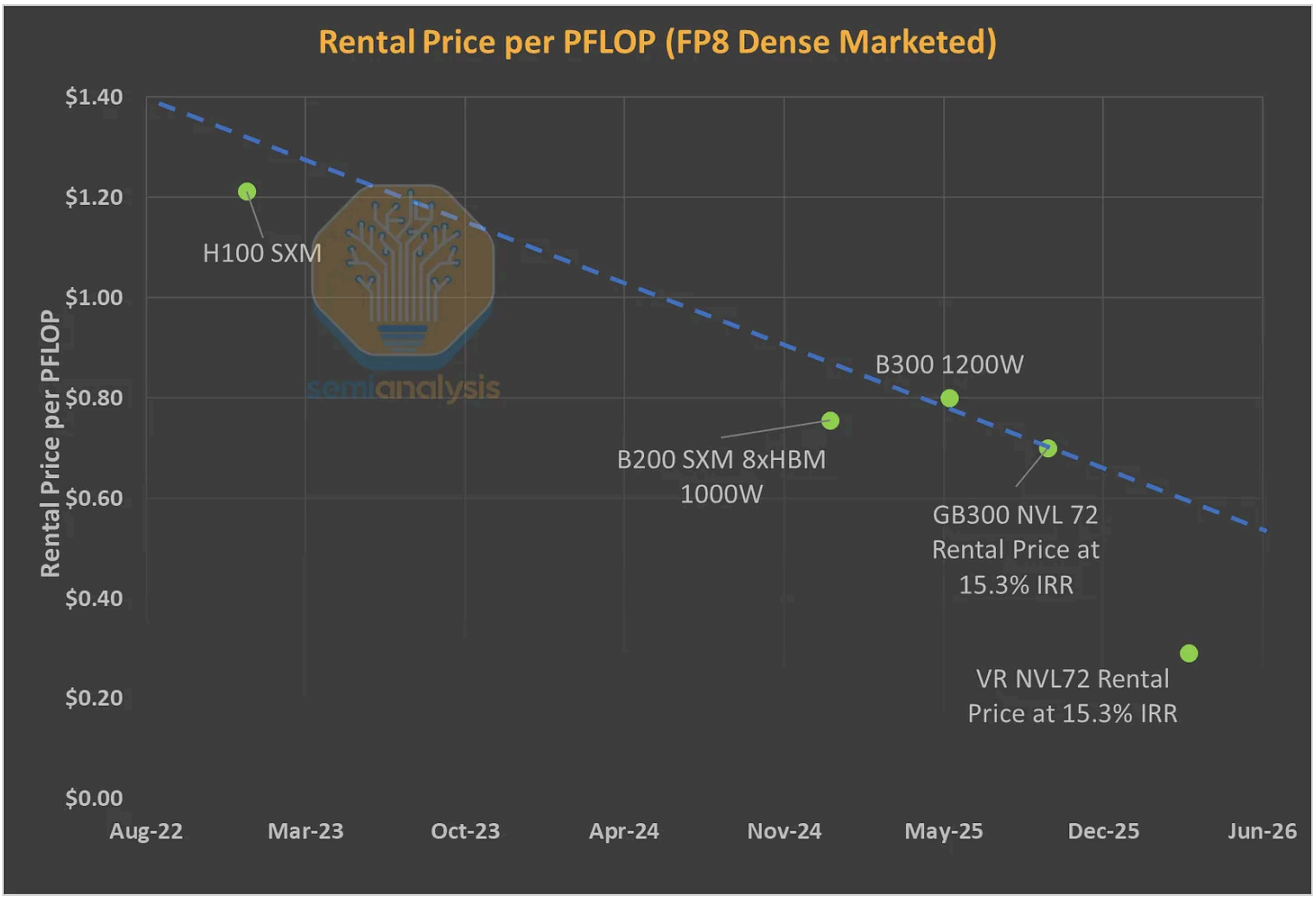


Source: SemiAnalysis AI TCO Model

SemiAnalysis AI TCO 模型

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.

我们的第二个框架着眼于基于价值的定价。我们以现有 SKU 所隐含的 $/FLOP 为基准，并探讨这对于 Rubin 意味着什么。这代表了计算租赁者在 Rubin 和当前一代 GPU 之间保持中立时愿意支付的理论最高价格——因此也充当了 GPU 租赁定价的天花板。在这里，我们关注每 PFLOP 租赁价格改善的趋势。



Source: SemiAnalysis AI TCO Model

SemiAnalysis AI TCO 模型

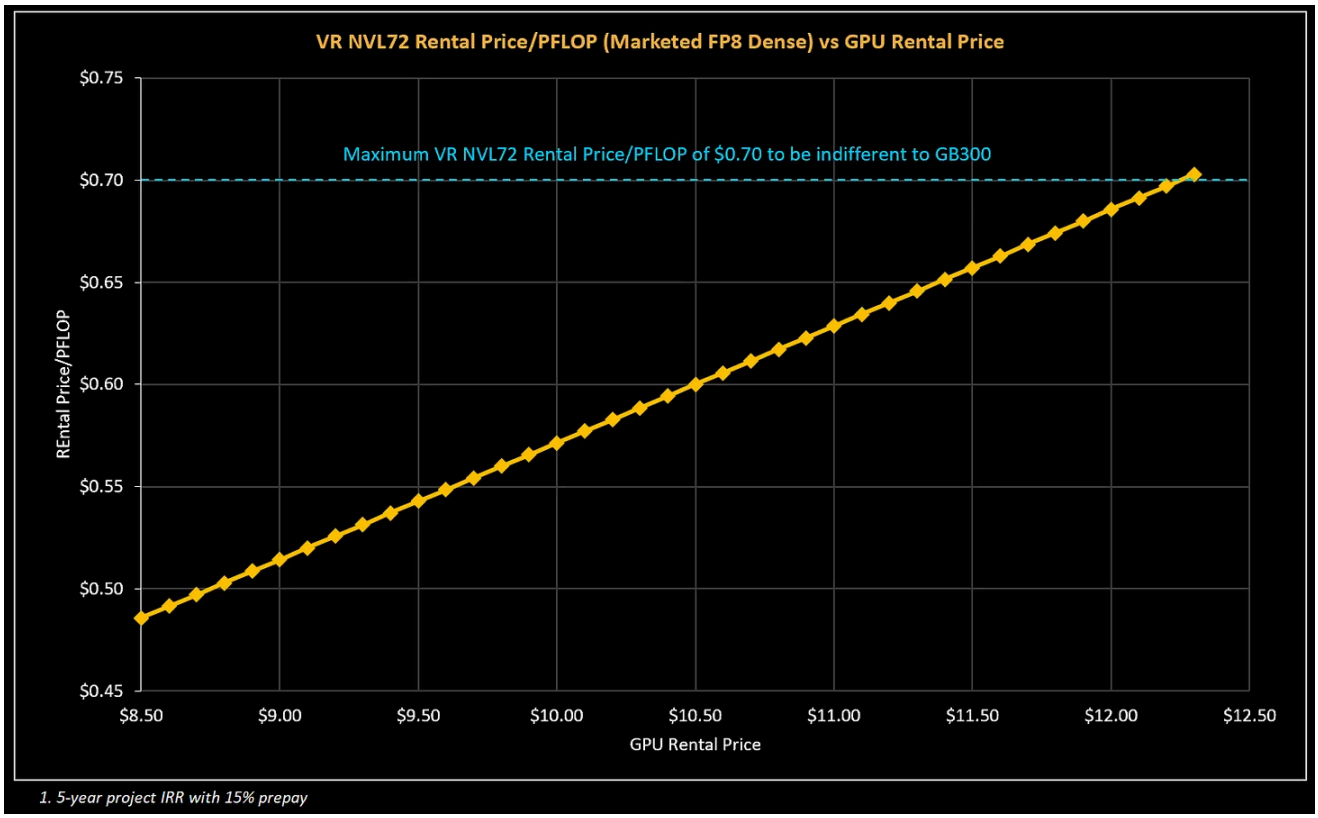
For training workloads, we anchor on GB300 pricing by comparing rental cost per PFLOP on a marketed FP8 dense basis. Using a current 5-year GB300 rental price of ~$0.70 per PFLOP, we derive a VR NVL72 ceiling price of approximately $12.25 per GPU hour at parity.

对于训练工作负载，我们以 GB300 的定价为基准，通过比较市售 FP8 稠密算力下每 PFLOP 的租赁成本来进行评估。按照目前 GB300 五年期租赁价格约每 PFLOP 0.70 美元计算，我们得出在同等性价比下，VR NVL72 的上限价格约为每 GPU 小时 12.25 美元。

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.

VR NVL72 的单位 TCO 价格表现尤为突出，因为与 GB300 及之前的显卡不同，其基于价值的定价与基于成本的定价之间存在极大的差距。如果我们保守地选择趋势线稍下方的一个点——例如每 PFLOP 0.55 美元的租赁价格——这将对应于 9.63 美元/小时/GPU，几乎是跨越 Neoclouds 回报门槛所需的最低租赁价格（4.92 美元/小时/GPU）的两倍。



Source: SemiAnalysis AI TCO Model

SemiAnalysis AI TCO 模型

One Chart to Rule Them All

一图胜千言

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.

基于成本的方法构成了 GPU 租赁定价的底线——如果低于这一租赁价格，Neoclouds 将不会批准新的 GPU 项目。基于价值的方法则构成了 GPU 租赁定价的理论上限——没有客户会为了租赁新一代 GPU 而支付更高的单位算力（$/FLOP）成本。

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.

我们将这两个约束条件以及一条说明在给定 GPU 租赁价格下 Neocloud 回报率的价格曲线结合起来，创建了一个定价图表。这张“统领全局的图表”同时也作为一个框架，用于理解竞争动态和定价权。

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

在文章开头，我们提出了一个问题：强劲的 AI 需求所带来的利益正流向谁？

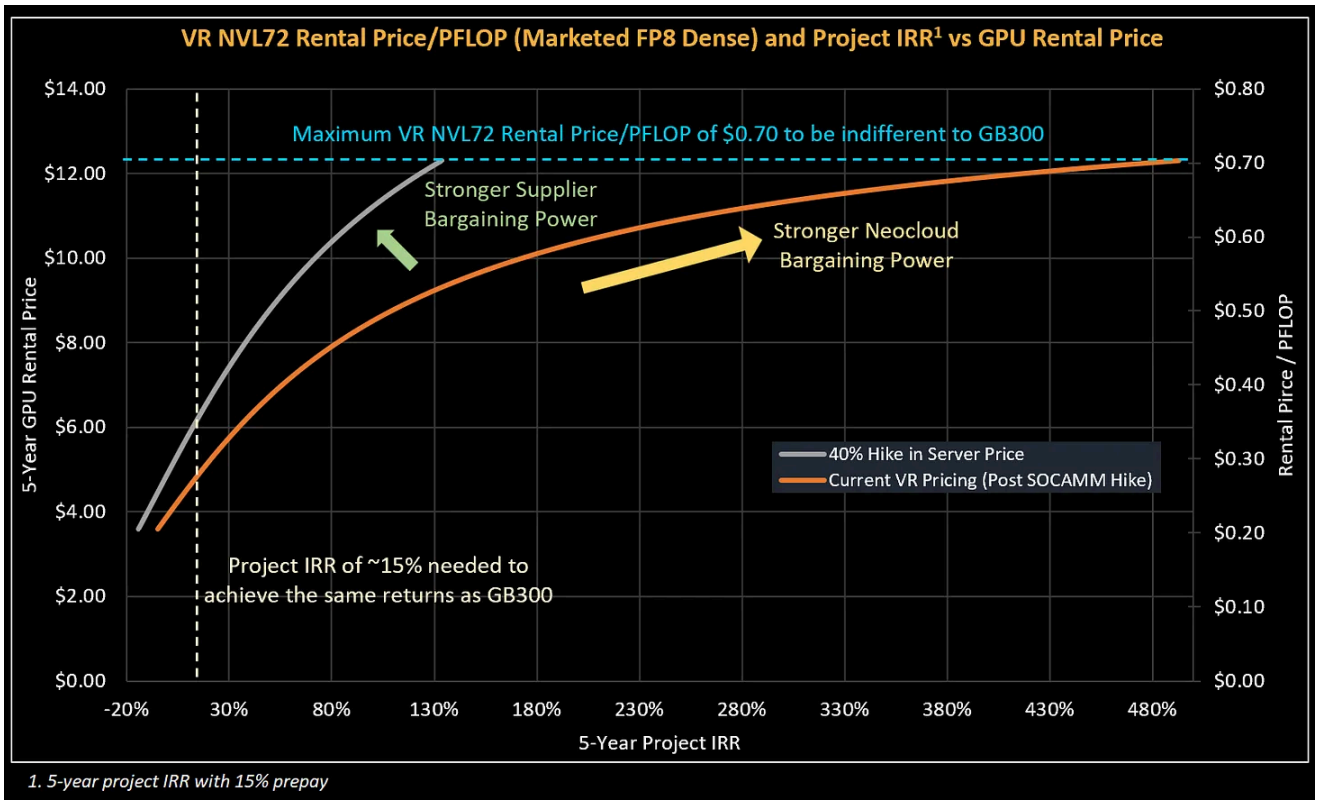
This question can be answered by plotting observed GPU Rental Prices charged by Neoclouds and the IRRs earned by these projects. Sliding up and to the right along the orange curve in the chart below represents stronger Neocloud bargaining power: The Neocloud is able to charge higher GPU Rentals and earn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

这个问题可以通过绘制 Neocloud 收取的 GPU 租赁价格观测值以及这些项目所获得的内部收益率（IRR）来回答。在下方图表中，沿着橙色曲线向右上方移动代表 Neocloud 的议价能力更强：Neocloud 能够收取更高的 GPU 租金，并获得远高于其 IRR 门槛率的收益。

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.

如果 Nvidia 提高 GB200 NVL72 的定价，定价曲线将向左上方移动。这是因为从 Neocloud 的角度来看，在保持内部收益率（IRR）不变的情况下，需要更高的租赁价格来抵消这一更高的系统成本。这种转变代表了像 Nvidia 这样的系统供应商拥有更强的议价能力。



Source: SemiAnalysis AI TCO Model

SemiAnalysis AI TCO 模型

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.

左上角是蓝色最高租赁价格/PFLOP 与米色 Neocloud 项目 IRR 最低门槛的交汇处，代表了 AI 集群的理论最高定价。如果系统定价高于此水平，Neocloud 和最终用户还不如直接购买或租赁 GB300。当前定价曲线与左上角之间的差距越大，意味着像 Nvidia 这样的 AI 集群提供商提高系统定价的空间就越大。

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.

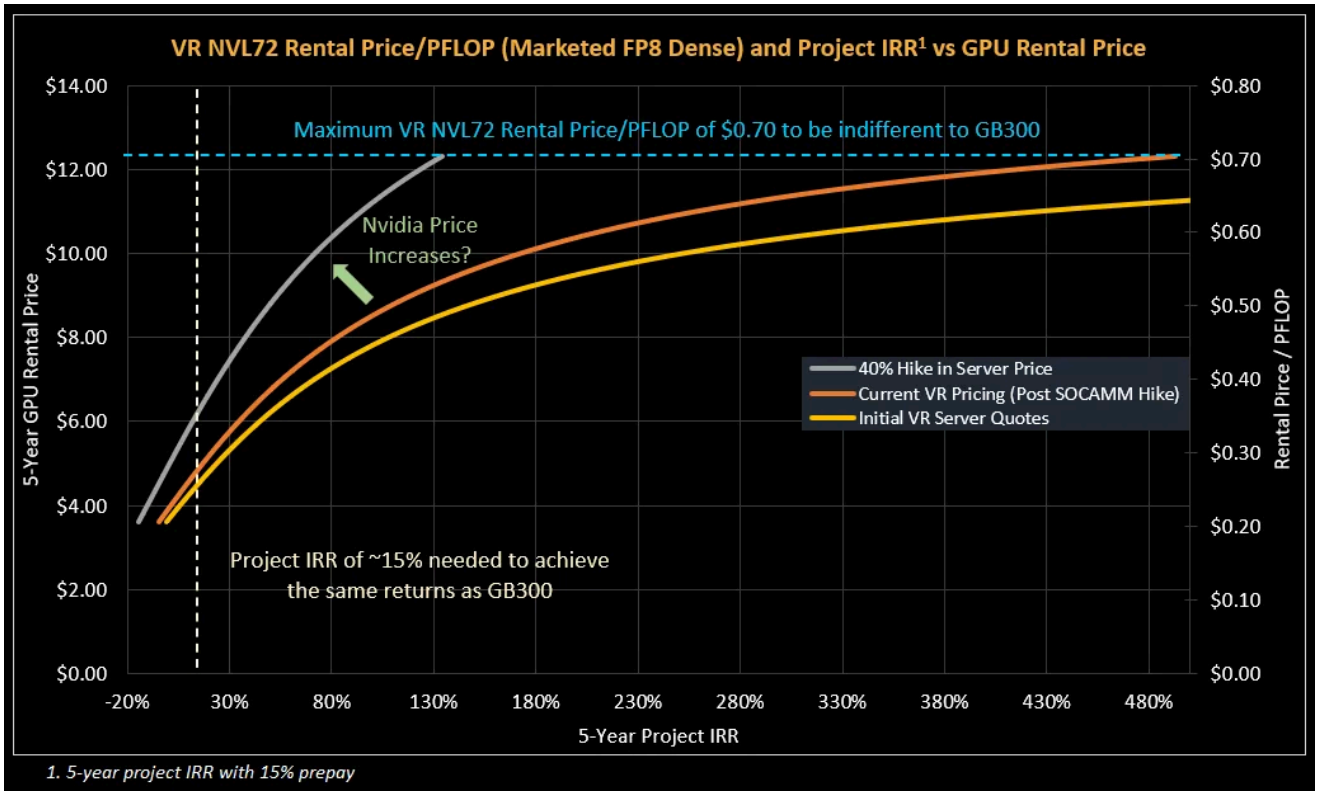
按照目前 VR NVL72 系统的定价，Neoclouds 在为期 5 年的合同中可以收取每 GPU 每小时 4.90 美元的费用，同时仍能获得与 GB300 项目相同的 15% 内部收益率（IRR）。对于客户而言，每 PFLOP 的租赁价格计算下来为 0.28 美元，与 GB300 NVL72 相比，每 PFLOP 的成本下降了 60%，这一成本改善幅度远低于历史趋势。

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.

这表明 Nvidia 在提高服务器价格方面仍有相当大的空间。服务器价格上涨约 40% 将导致每 FLOP 价格的成本改善低于趋势水平，但仍能为 Neoclouds 留出足够的提价空间，从而获得更高的 IRR。即使 Neoclouds 通过提高定价来沿灰色曲线移动，例如每 GPU 每小时收费 8.00 美元并获得 38% 的 IRR，这对应于 0.46 美元/PFLOP 的成本，这仍然是一个低于趋势水平的改善。

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.

需要指出的是，这一分析主要集中在单位算力（FLOP）的租赁价格上，但单位总拥有成本（TCO）的推理性能提升速度甚至更快。尽管我们尚未在 InferenceX 上对 VR NVL72 系统进行基准测试，但极有可能的是，VR NVL72 在单位 Token 交付成本方面的下降速度会更加剧烈，这意味着 Nvidia 从整个生态系统中获取更多价值的空间可能还会更大。



Source: SemiAnalysis AI TCO Model

SemiAnalysis AI TCO 模型

VR NVL72 vs GB300 Performance per TCO

VR NVL72 与 GB300 的单位总拥有成本（TCO）性能对比

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

Rubin 在绝对性能上实现了显著提升，但更重要的问题在于，这种提升如何转化为单位 TCO（总拥有成本）下的性能表现。

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.

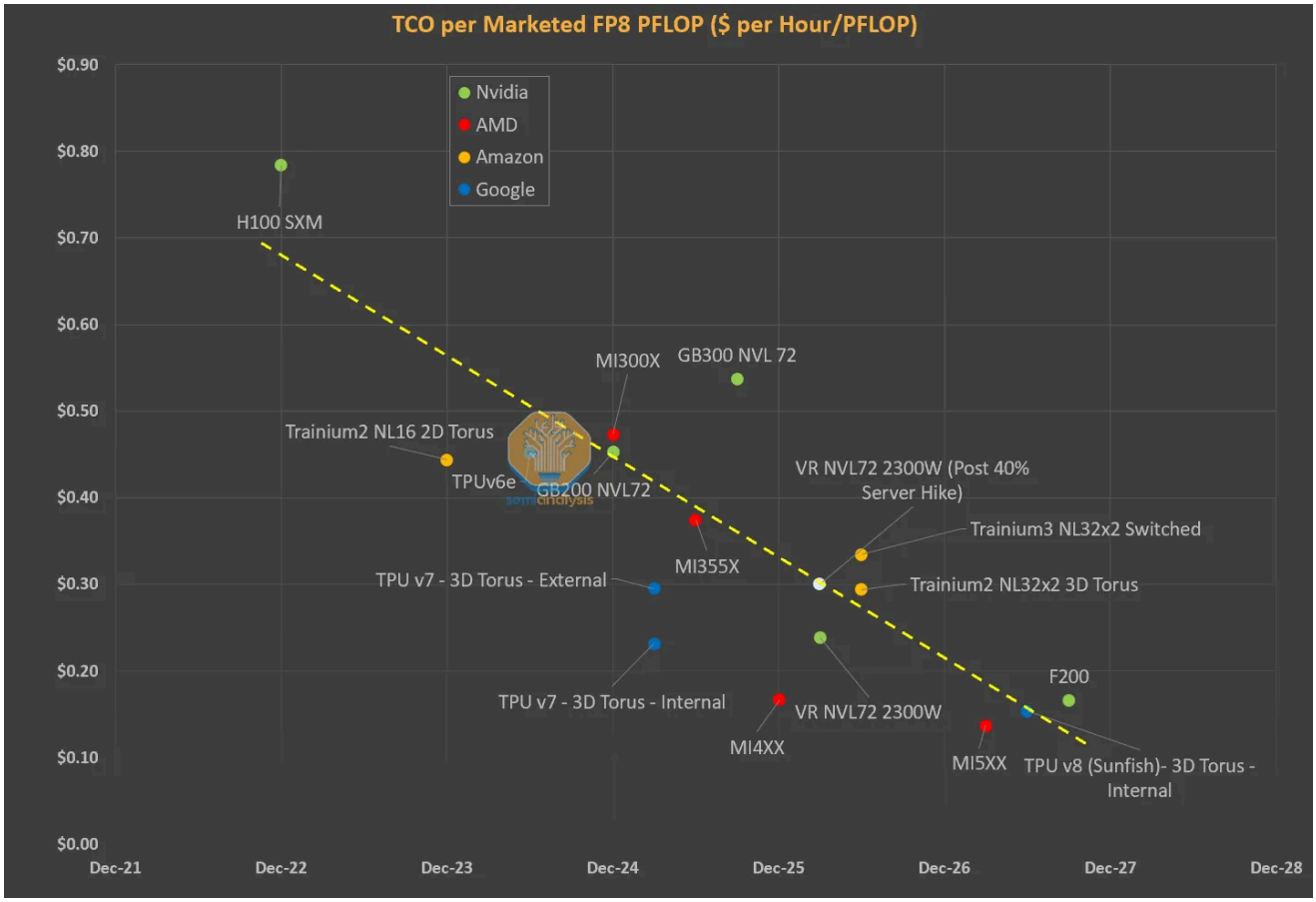
在系统层面，VR NVL72 具有更高的成本基数。每 GPU 每小时的总成本从 GB300 NVL72 的 2.69 美元增加到 VR NVL72 的 4.18 美元，这反映了更高的系统成本，而这在很大程度上是由内存价格上涨驱动的。

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.

宣传的稠密 BF16 性能从 2,500 TFLOPS 提升至 4,000 TFLOPS，而宣传的稠密 FP8 性能则从 5,000 TFLOPS 扩展至 17,500 TFLOPS。最大的跨越来自 FP4，其稠密性能从 15,000 TFLOPS 飙升至 35,000 TFLOPS。显存带宽也显著增加，从每个逻辑 GPU 的 8 TB/s 提升至 22 TB/s。Nvidia 还在宣传 FP4 稀疏性，给出的数据是 50,000 TFLOPS（对比稠密性能的 35,000 TFLOPS）。如果这能被有效利用，将进一步降低单位性能成本，使每 PFLOP 的 TCO 从 0.12 美元降至 0.08 美元，降幅约 33%。这是一个重要的杠杆，但前提是取决于实际工作负载的兼容性。

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.

从总拥有成本（TCO）的角度来看，Rubin 在 BF16 精度上的提升较为有限，每小时每 PFLOP 的成本从 1.07 美元略微下降至 1.04 美元。而在更低精度下，这种提升变得更加显著。FP8 的单位 PFLOP TCO 从 0.54 美元降至 0.24 美元，而 FP4 稠密计算则从 0.18 美元优化至 0.12 美元。这些进步反映了架构向低精度计算的转型，以及 Nvidia 对工作负载将日益向低精度栈迁移的预期。



Source: SemiAnalysis AI TCO Model

来源：SemiAnalysis AI TCO 模型

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.

这些对比是基于市场宣传性能而非有效吞吐量。实际性能取决于模型算力利用率（MFU），而 MFU 会因工作负载和对系统的熟悉程度而产生显著差异。这也是 Rubin 在首次部署时实现基于价值定价的另一个制约因素，因为在软件支持和工程经验成熟之前，初始部署时的 MFU% 可能会较低。

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.

Compute Competition 算力竞争

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).

盈利能力最终由竞争决定。当竞争激烈时，定价会趋向于成本（即较低的利润率）；而垄断者则可以根据价值或客户所能承受的最大限度进行定价（较高的利润率）。

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.

Nvidia 正在向市场供应稀缺的算力，而像 Anthropic 这样的客户正在产生强劲的推理利润，这表明他们对算力的付费意愿显著高于当前的定价水平。

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.

然而，这忽略了英伟达在计算层面的竞争。去年的一个关键故事是 Anthropic 如何成功实现了计算资源的多样化，摆脱了对英伟达系统的依赖。Mythos 并非在英伟达设备上训练。长期以来，除了谷歌之外，英伟达几乎是所有公司首选的现货 GPU 供应商。

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.

首先，Anthropic 大幅转向了 Trainium 和 TPU，我们在之前的文章中详细报道了这一进展。虽然 TPU 和 Trainium 在硬件和软件方面并不具备对英伟达 GPU 的绝对优势，但这一短板被更低的成本所弥补。亚马逊和谷歌支付给 Marvell、Alchip、Broadcom 和 Mediatek 等设计合作伙伴的利润率，低于英伟达向其客户收取的费用，并且它们仍能通过转售算力来降低单位 Token 成本或单位训练 FLOP 成本。有人认为，正是这种竞争引力抑制了英伟达利润率的进一步上升。

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.

即使英伟达保持着明显的性能领先地位，可靠的低成本替代方案的存在，也限制了其在不加速客户流失（转向多样化方案）的前提下激进提价的空间。

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

最终，问题在于英伟达是否能够且愿意将其性能优势转化为定价权。

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

我们的结论是，他们应该趁热打铁，充分利用其在内存定价、容量和性能方面的长期优势。

Recommend 推荐


71 Likes 71 次赞 · 8 Restacks 8 次转推

Previous 上一篇


A guest post by
Crystal Huang

Subscribe to Crystal

Discussion about this post

关于此帖的讨论

Comments 评论 Restacks 转推


Write a comment...