Semis Memo: Supply Chain Inheritance

半导体备忘录：供应链传承

Power & Analog Semis, CPUs in the Agentic Era, Neoclouds, Material Bottlenecks, Korea Unlocked

功率与模拟半导体、智能体时代的 CPU、新云服务商、材料瓶颈、韩国市场解锁

CITRINI

MAY 12, 2026 2026 年 5 月 12 日 · PAID · 已付费

157 1 4 Share 分享

Introduction 引言

For the first innings of the AI Infrastructure trade, it was simple enough just to know the basics. Large Language Models run on GPUs, buy Nvidia. AI compute will lift optics out of the telecom doghouse and cause significant growth for interconnects, buy the optical interconnect names. Every iota of AI compute demand in the agentic era must inevitably flow through the memory OEMs, buy Micron and SK Hynix.

在 AI 基础设施交易的第一阶段，只需了解基本常识就足够了。大语言模型在 GPU 上运行，那就买 Nvidia。AI 计算将把光模块从电信行业的低谷中拯救出来，并带动互连器件的显著增长，那就买光互连相关的股票。在智能体时代，每一分 AI 计算需求最终都必然流向存储 OEM 厂商，那就买美光（Micron）和 SK 海力士（SK Hynix）。

It seemed a lot more difficult than that at the time, but it was pretty much that simple. The reasons for that were twofold. First, not everyone bought into the massive growth in data centers that would be required for AI to proliferate and actualize. Second, between the post-COVID supply chain glut and numerous other headwinds to semiconductor and adjacent names, valuations remained quite forgiving in all but the

most obvious first-order beneficiaries.

虽然当时看起来比这复杂得多，但实际上基本就是这么简单。原因有两点：首先，并非所有人都相信 AI 的普及和实现需要数据中心如此大规模的增长。其次，在后疫情时代的供应链过剩以及半导体及相关领域面临的众多其他不利因素影响下，除了那些最显而易见的一阶受益者外，估值水平依然相当宽容。

That's begun to change, and with it, outperforming in the AI infrastructure complex requires in-depth understanding beyond just identifying current bottlenecks. Understanding the roadmap for the future requires a bit more technical competence, which is why in January 2026 we **began** our Semis Memo series – guided by our semis analysts Zephyr and Jukan.

这种情况已经开始发生变化，随之而来的是，要在 AI 基础设施领域获得超额收益，不仅需要识别当前的瓶颈，还需要更深入的理解。了解未来的路线图需要更强的技术能力，这也是为什么我们在 2024 年 1 月开启了由半导体分析师 Zephyr 和 Jukan 指导的“半导体备忘录”（Semis Memo）系列。

While the landscape has evolved, our framework stays the same. Begin with the macro. Find areas where forecasts are still reflecting overhangs from non-AI related headwinds and determine whether AI demand can overcome them in a way that makes estimates too low.

虽然格局在不断演变，但我们的框架保持不变。首先从宏观入手，寻找那些预测仍反映出非 AI 相关不利因素影响的领域，并判断 AI 需求是否能以某种方式克服这些影响，从而导致当前的预期估值过低。

In this issue, we're covering the following places that meet our criteria:

在本期内容中，我们将涵盖符合我们标准的以下领域：

● **Analog and Power Semis: Supply Chain Inheritance**

● 模拟与功率半导体：供应链的传承

● **CPUs in the Agentic Era**

● 智能体时代的 CPU

● **Neoclouds: The Inference Shortage**

● 新型云服务商：推理算力短缺

● **AI Materials Bottlenecks**

● AI 材料瓶颈

● **Korea Unlocked** ● 韩国市场的潜力释放

● **Updating Previous Semis Memo Ideas**

● 更新之前的半导体备忘录观点

We end with Some Thoughts on Where We're Going...

最后，我们对未来的发展方向进行了一些思考.....

Analog and Power Semis 模拟与功率半导体

We first flagged the likelihood that AI demand would overwhelm the headwinds currently being experienced by the analog and power semi sector in our **25 Trades for 2025**, specifically as it related to the upcoming Multilayer Ceramic Capacitors (MLCC) shortage.

在我们的《2025 年的 25 笔交易》中，我们首先指出了 AI 需求可能会抵消模拟和功率半导体行业目前面临的逆风，特别是涉及到即将到来的多层陶瓷电容器（MLCC）短缺问题。

> *Components integral to power quality management systems address common issues such as voltage sags, harmonics, and transients, thereby ensuring the reliable operation of electrical and electronic equipment. This includes capacitors, inductors, diodes, power ICs, surge*

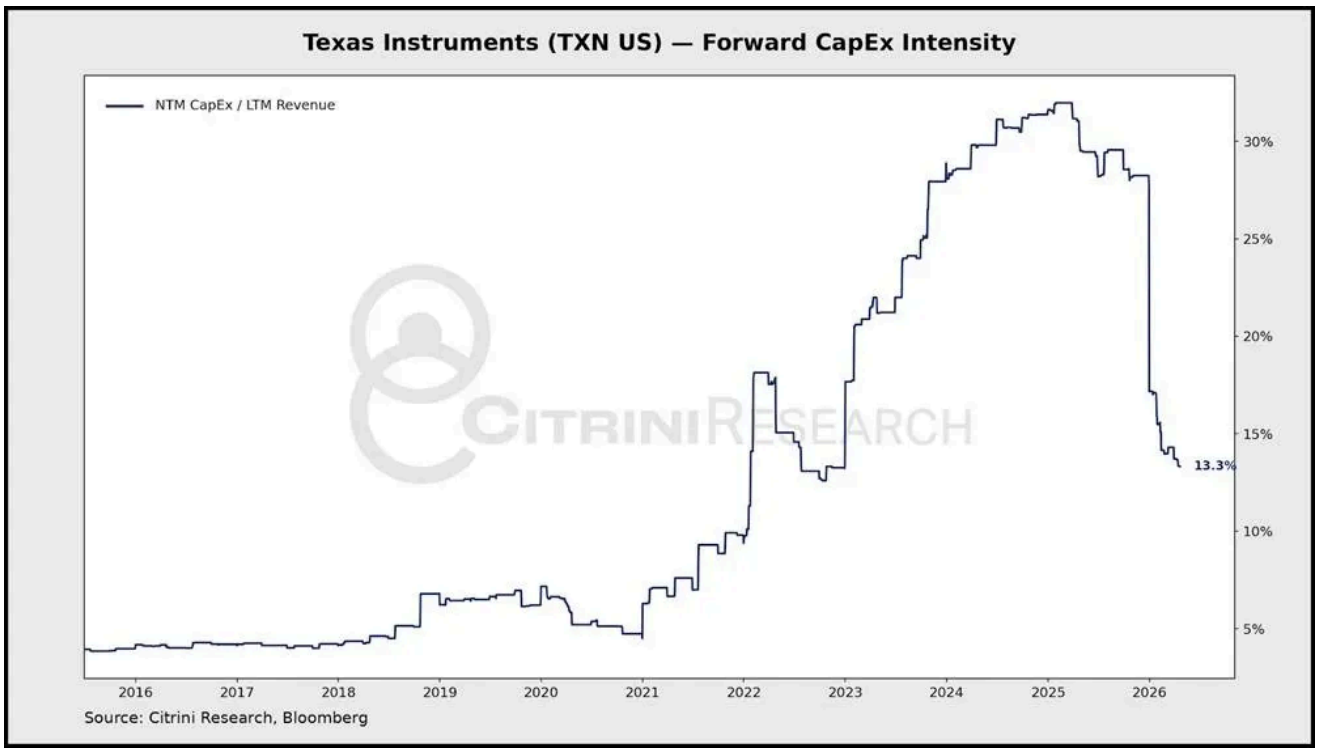
> *protectors, filters, transformers, uninterruptible power supplies (UPS).*
>
> *电能质量管理系统不可或缺的组件解决了电压骤降、谐波和瞬态等常见问题，从而确保了电气和电子设备的可靠运行。这包括电容器、电感器、二极管、电源IC、浪涌保护器、滤波器、变压器、不间断电源（UPS）。*
>
> *Discrete power semiconductors (like MOSFETs and diodes) will also benefit as they are integral to creating efficient, stable power rails. Filters, ferrite beads, and connectors may see growth, but* ***the clearest secular uplift is likely in capacitors and inductors*** *given their centrality to power conversion in AI-driven, high performance computing environments.*
>
> *分立功率半导体（如MOSFET和二极管）也将从中受益，因为它们是构建高效、稳定电源轨不可或缺的组成部分。滤波器、磁珠和连接器可能会迎来增长，但考虑到电容器和电感器在AI驱动的高性能计算环境功率转换中的核心地位，它们很可能获得最明确的长期增长动力。*

These names have begun to outperform, and we feel it's directly related to another framework we've posed for 2026 – "**Post-Traumatic Supply Disorder**". The companies dealing with power semis have had to contend with a barrage of headwinds – the COVID supply glut, competition from Chinese analog semis, the anemic EV and automotive cycle…the list goes on. However, they're beginning to see data center revenues climb. And they're not rushing to add capacity, having been burnt one too many times.

这些标的已经开始表现出超额收益，我们认为这与我们为2026年提出的另一个框架——"供应后创伤障碍"（Post-Traumatic Supply Disorder）直接相关。经营功率半导体的公司此前不得不应对一系列不利因素——新冠疫情后的供应过剩、来自中国模拟芯片的竞争、疲软的电动汽车和汽车周期……名单还在继续。然而，他们正开始看到数据中心业务的收入攀升。而且，由于被"烫"过太多次，他们并不急于增加产能。

Take a look at the capex intensity (capex / revenue) for **Texas Instruments** (**TXN US**):

看看德州仪器 (Texas Instruments, TXN US) 的资本支出强度（资本支出 / 营收）：



It's typically this part of the cycle that results in supply ramping, but instead, TXN and peers like **NXP Semiconductors** (**NXPI US**) are content to raise prices.

通常在周期的这个阶段，供应量会开始攀升，但相反，TXN 及其同行如 NXP Semiconductors (NXPI US) 却满足于提高价格。



Texas Instruments

May 7, 2026

Dear valued partner,

We are writing to inform you of upcoming adjustments to our pricing. New prices will apply to all orders and shipments effective July 1, 2026. We are sharing this notice with you well in advance, to be transparent with our plans and to help ensure prices are aligned for upcoming orders and payments.

These price adjustments will impact products across our portfolio, and the increases will be dependent upon the specific materials and technologies. These increases are driven by the current market environment and increasing costs throughout the broader supply chain.

We value your business and remain committed to providing you with competitive products, support, pricing, and availability. Thank you for the continued collaboration.

Regards,

Mark Roberts
Senior Vice President
Worldwide Sales and Marketing
Texas Instruments



NXP

May 1st, 2026

Subject: Notification of Price Changes

Dear Valued Customer,

We are writing to inform you of upcoming pricing changes. These adjustments are a direct result of evolving market conditions that continue to drive inflationary cost pressures across several key areas, including raw materials, energy, labor, logistics, and supplier inputs.

To address these increased costs, beyond our control, NXP will implement price adjustments **effective June 1st**.

NXP remains committed to investing in and delivering innovative product solutions that enable the success of your business.

As a next step, your NXP account manager will contact you to communicate part-specific pricing details and answer any questions.

Thank you again for your continued partnership.

Regards,

NXP Semiconductors

Now, we're at an inflection point, and these companies are letting ASPs go up rather than flood the market. Up until now, however, we've been mostly focused on the rack-

internal story. Companies like **Murata Manufacturing** (**6981 JP**), **Vishay Intertechnology** (**VSH US**) and **Samsung Electro-Mechanics** (**009150 KS**) have taken off as the crowd recognizes exactly how short on MLCCs we are.

现在，我们正处于一个转折点，这些公司宁愿让平均售价（ASP）上涨，也不愿向市场大量倾销。然而，到目前为止，我们主要关注的是机架内部的情况。随着大众意识到多层陶瓷电容器（MLCC）的短缺程度，村田制作所（Murata Manufacturing，6981 JP）、威世科技（Vishay Intertechnology，VSH US）和三星电机（Samsung Electro-Mechanics，009150 KS）等公司的股价已经开始腾飞。

For our first and highest-conviction section, we're glad to say that you don't have to be a semiconductor expert to understand the rack-external power semis story. It's a pretty cut and dry setup. **The capex that burned them once was actually the exact infrastructure necessary for this part of the cycle.**

对于我们第一个也是信心最强的板块，我们很高兴地告诉大家，你不需要成为半导体专家也能理解机架外功率半导体的故事。这是一个非常清晰明了的局面。曾经让他们蒙受损失的资本支出，实际上正是这一周期阶段所必需的基础设施。

While we've long been waiting for the automotive overhang to lift the fog off of the names in the analog and power semis space, we're now realizing something more significant. It doesn't really need to – rather, the AI capex buildout is simply inheriting the EV buildout supply chain.

Supply Chain Inheritance 供应链传承

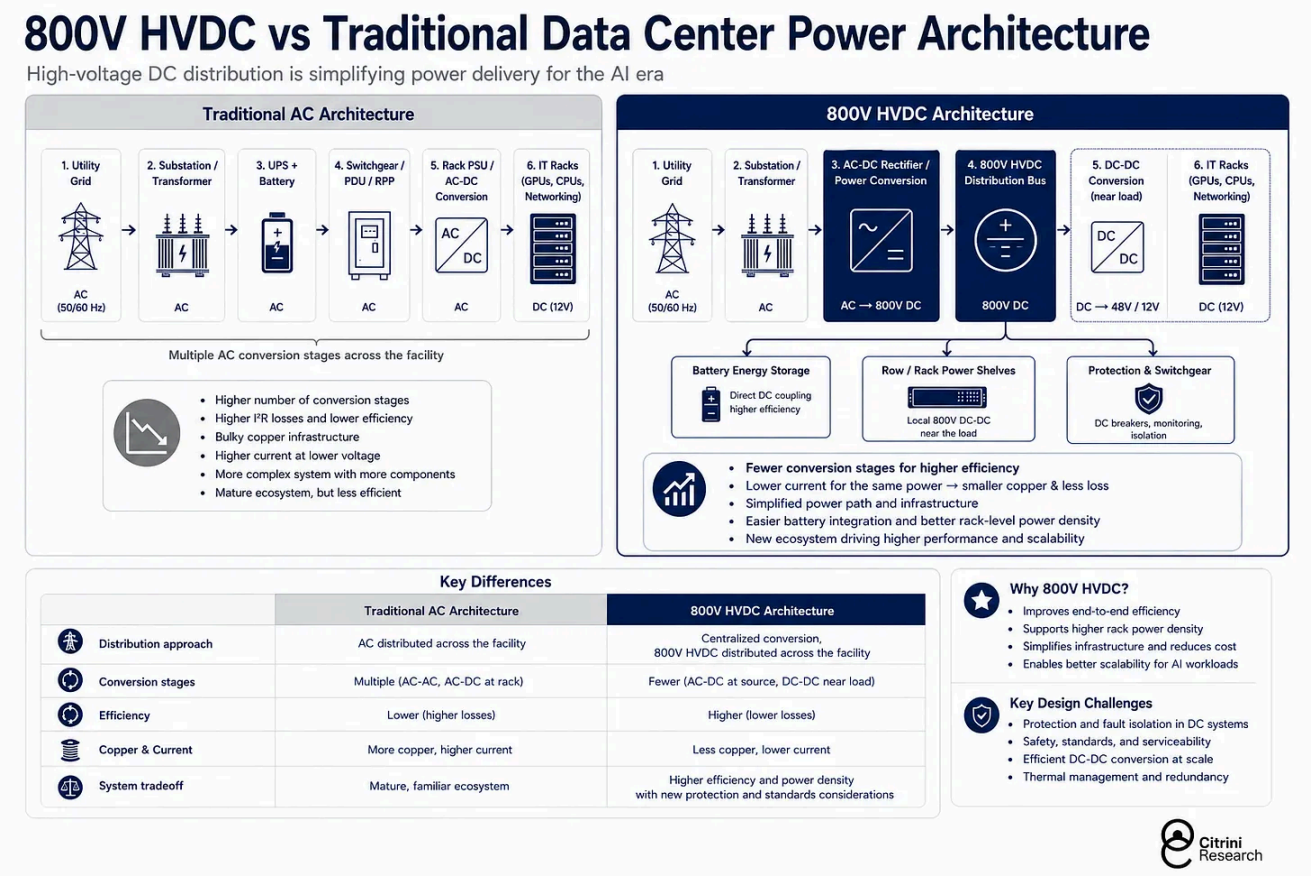
In Nvidia's May 2025 technical blog on 800V DC rack architecture, they credit the underlying technology to "the electric vehicle and solar industries." *That's the trade...*

在英伟达 2025 年 5 月关于 800V 直流机架架构的技术博客中，他们将底层技术归功于“电动汽车和太阳能行业”。这就是交易的核心所在……

The most consequential power architecture transition in the history of computing is happening because Porsche, BYD, and a handful of solar inverter companies spent the last decade scaling a supply chain for completely unrelated end markets. Nvidia inherited both the silicon and a half-decade of cost-curve compounding – walking into

an automotive-validated, fully ramped wide-bandgap semiconductor industry.

计算历史上影响最深远的电力架构转型之所以正在发生，是因为保时捷、比亚迪和少数几家太阳能逆变器公司在过去十年中，为一个完全不相关的终端市场规模化了供应链。英伟达继承了这些硅技术以及五年的成本曲线复利效应——直接步入了一个经过汽车行业验证、产能充足的宽禁带半导体产业。



We've spoken about this before (in our **2025 Robotics Update**), but it seems that the market is becoming increasingly aware of it now. Anyone who spent 2025 dumping SiC names on soft European EV demand is about to find out those names have a second customer cohort showing up at exactly the moment the first one is wobbling. Now, companies that spent the last 5 years in the doldrums of the EV and COVID supply glut aftermath find themselves with a massive moat made up of the capex

they've spent *(some of whom, the capex that bankrupted them! More on that later)*.

我们之前讨论过这一点（在我们的《2025 年机器人行业更新》中），但市场现在似乎正日益意识到这一点。任何在 2025 年因欧洲电动汽车需求疲软而抛售碳化硅（SiC）相关股票的人，很快就会发现，这些公司正迎来第二批客户群体，而这恰好发生在第一批客户动摇的时刻。现在，那些在过去五年里深陷电动汽车和新冠疫情供应过剩后遗症泥潭的公司，发现自己拥有了一道由已投入资本支出构成的巨大护城河（其中一些公司的资本支出甚至导致了它们的破产！稍后会详细介绍）。

This supply chain inheritance is a result of AI capex becoming the apex predator of the global economy, sucking capital and capacity out of every adjacent industry. Here, AI sits on the receiving end of capacity that was built for someone else, with no requirement that the original industry stay healthy for the trade to work. (Indeed, the trade works *better* if the original industry is currently struggling, because the names are already mispriced on the original cohort's weakness.)

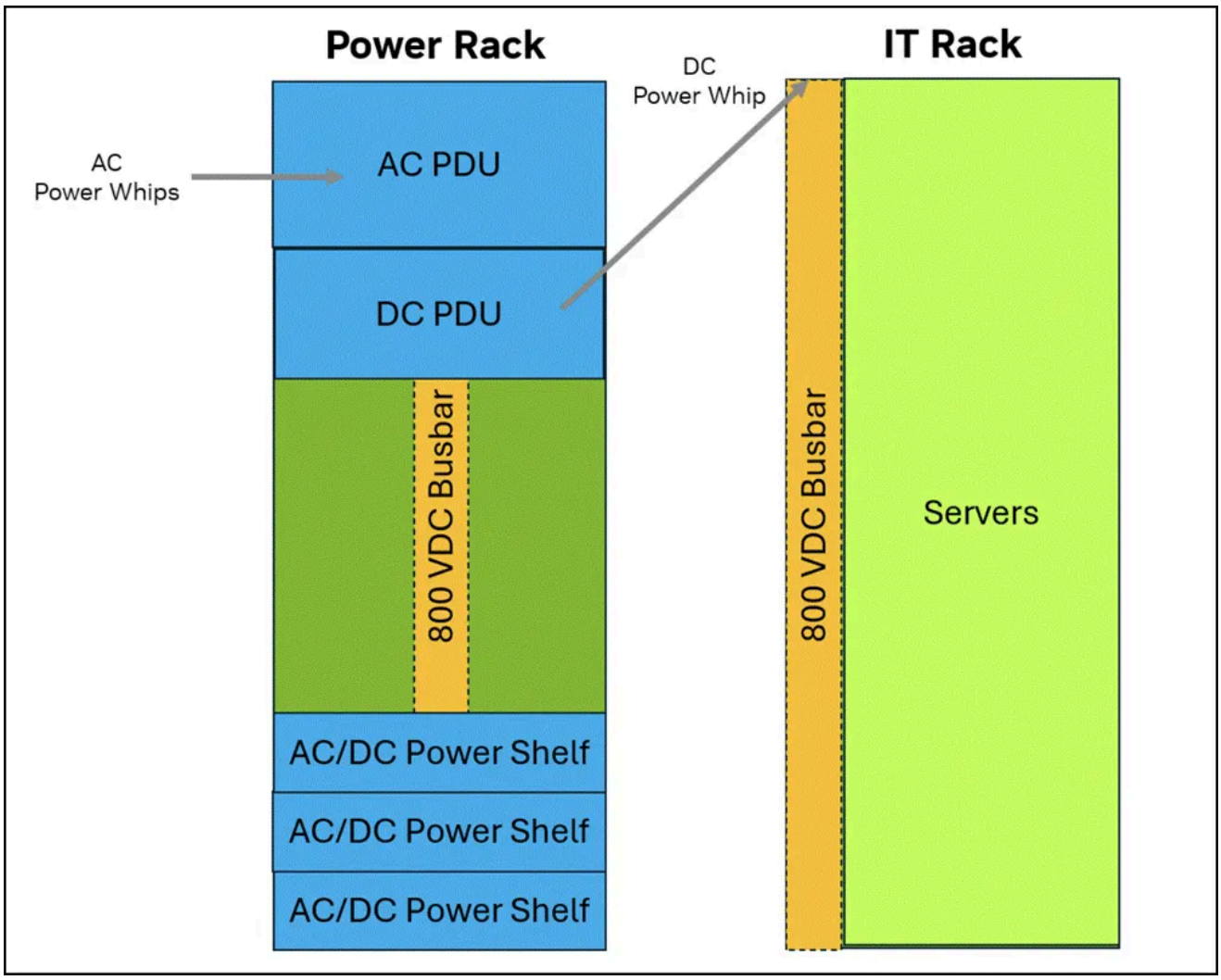
这种供应链继承是人工智能资本支出成为全球经济“顶级掠食者”的结果，它从每一个相邻行业中吸取资金和产能。在这种情况下，人工智能处于接收端，承接了原本为其他行业建设的产能，且这种交易的运作并不要求原有的行业保持健康。（事实上，如果原有的行业目前正处于困境，这种交易的效果反而更好，因为这些标的已经因原有行业的疲软而被错误定价了。）

The Porsche engineers who specified the Taycan's 800V powertrain in 2017 had no idea they were laying the foundation for a 600kW GPU rack in 2027. *They were, though.*

2017 年，保时捷工程师在为 Taycan 设定 800V 动力总成规格时，绝不会想到他们正在为 2027 年的 600kW GPU 机架奠定基础。但事实的确如此。

This is the next layer of the AI Power Plumbing thesis we've been building all year. The MLCC and discrete power semi work was the rack-internal story. The Inheritance is the rack-external story.

这是我们全年构建的“AI 电力管道”论题的下一层级。关于 MLCC 和分立功率半导体的研究是机架内部的故事。而“传承”则是机架外部的故事。



The architecture itself is simpler than Nvidia's engineering blog makes it look. *(As always with Nvidia, the technical post goes hard on the physics, but the trade implications are direct.)* Voltage doubles from the legacy 54V rack standard to 800V, which cuts copper requirements substantially and, more importantly, is **the only path to physically building 600kW racks**.

该架构本身比 Nvidia 工程博客上描述的要简单。（一如既往，Nvidia 的技术文章在物理原理上着墨颇多，但其对贸易的影响却是直接的。）电压从传统的 54V 机架标准翻倍至 800V，这大幅减少了对铜的需求，更重要的是，这是物理上构建 600kW 机架的唯一途径。

The conversion stages from utility power down to the chip need wide-bandgap silicon at every step: **silicon carbide** (**SiC**) for the front-end down-conversion to 800V DC, **gallium nitride** (**GaN**) for the high-frequency step down to the intermediate voltages feeding the GPU. Both have been in volume production for half a decade because of

industries (EVs, fast-charging, solar inverters) that have nothing to do with AI.

从市电到芯片的转换阶段，每一步都需要宽禁带半导体：碳化硅（SiC）用于前端降压至 800V 直流电，氮化镓（GaN）用于高频降压至供给 GPU 的中间电压。由于与 AI 无关的行业（电动汽车、快速充电、太阳能逆变器）的推动，这两者投入量产已有五年之久。

Do you know what it looks like when a stock goes from *"beta to anemic EV/solar industry plagued by Chinese oversupply"* to *"critical supplier of the components necessary to jump over the power hurdle"*?

你是否见过这样一只股票：它最初只是"受中国产能过剩困扰、估值低迷的电动汽车/太阳能行业贝塔股"，随后却摇身一变，成为了"跨越电力障碍所需核心组件的关键供应商"？

Neither do I, but this is probably closer to the start of it than the end.

我也没见过，但这可能更接近它的开始，而非结束。



Source: Citrini Research

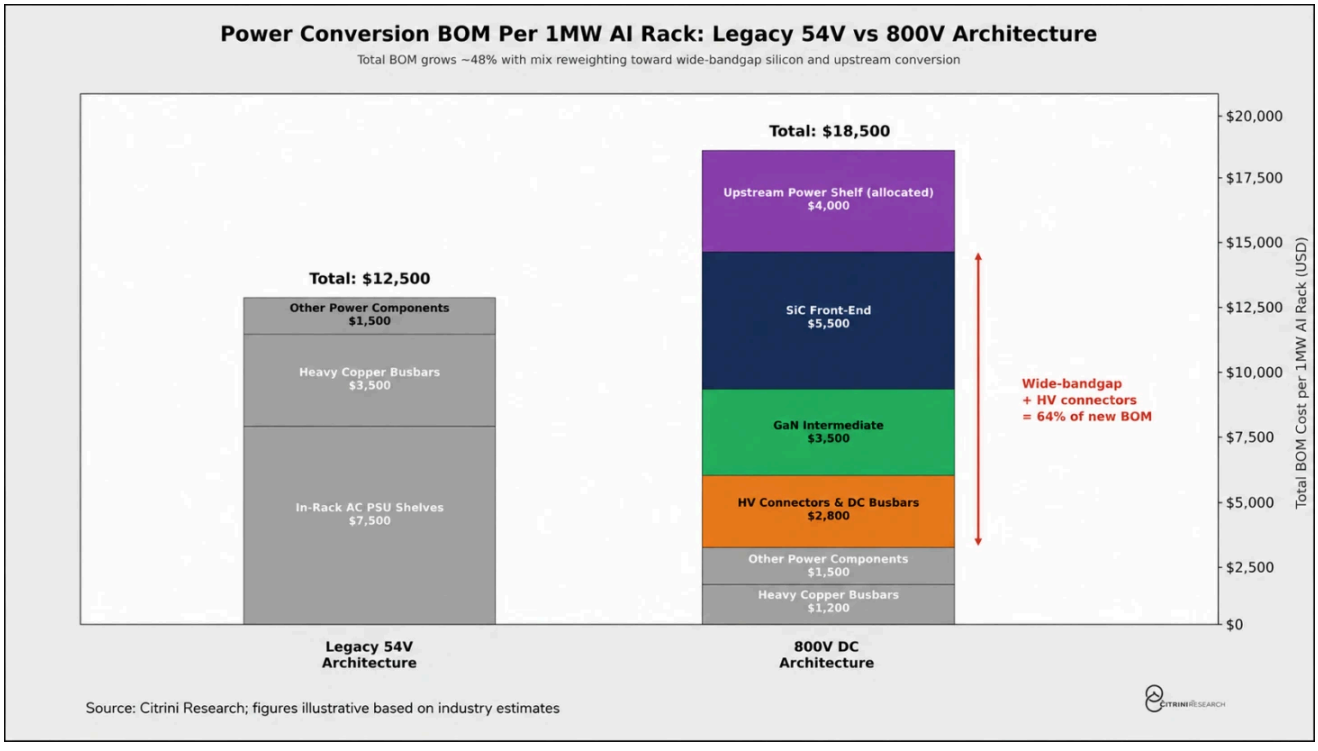
Today, eight power shelves per GB200 NVL72 sit inside the rack doing the conversion. At megawatt-class density those shelves would eat the entire rack, leaving zero room for compute. In the new architecture, conversion moves upstream to a dedicated

power shelf (sometimes called a sidecar), the IT rack receives DC directly.

如今，每个 GB200 NVL72 机架内配有八个电源层架（power shelves）负责电力转换。在兆瓦级密度下，这些层架将占据整个机架，导致没有空间放置计算单元。在新架构中，转换过程移至上游的专用电源层架（有时称为 sidecar），IT 机架则直接接收直流电（DC）。

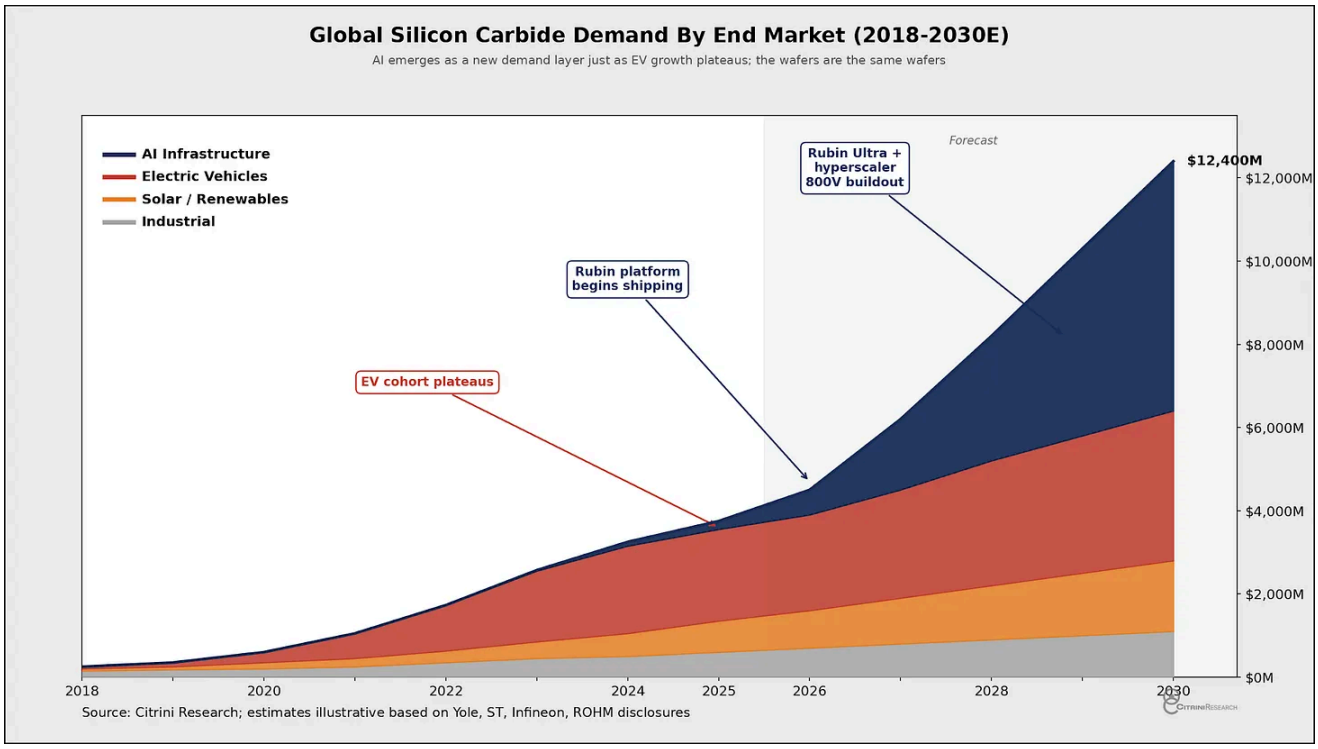
The picture below shows the bill-of-materials reweighting at the rack level. Total power BOM grows roughly 16% on the architecture transition (which is itself a tailwind for the supply chain in aggregate), but the mix shift is the actual story. Wide-bandgap silicon plus high-voltage connectors goes from zero percent of rack power BOM to roughly 64%.

下图显示了机架层级的物料清单（BOM）权重重组。随着架构转型，总电源 BOM 大约增长了 16%（这本身对整个供应链来说是一个利好），但真正的核心在于结构性的转变。宽禁带半导体加上高压连接器在机架电源 BOM 中的占比从零增长到了大约 64%。



The chart below shows global silicon carbide demand by end market from 2018 through our 2030 estimate.

下图显示了从 2018 年到我们预测的 2030 年，按终端市场划分的全球碳化硅需求情况。



The EV cohort drove SiC from a $300 million industrial niche in 2018 to a $3.5 billion market by 2024, a roughly 12x expansion in six years. That growth has been decelerating into a plateau as European EV demand softens and Chinese capacity takes share.

电动汽车群体推动碳化硅（SiC）从 2018 年仅 3 亿美元的工业利基市场，增长到 2024 年的 35 亿美元市场，六年间扩张了约 12 倍。随着欧洲电动汽车需求放缓以及中国产能抢占市场份额，这种增长已开始减速并进入平台期。

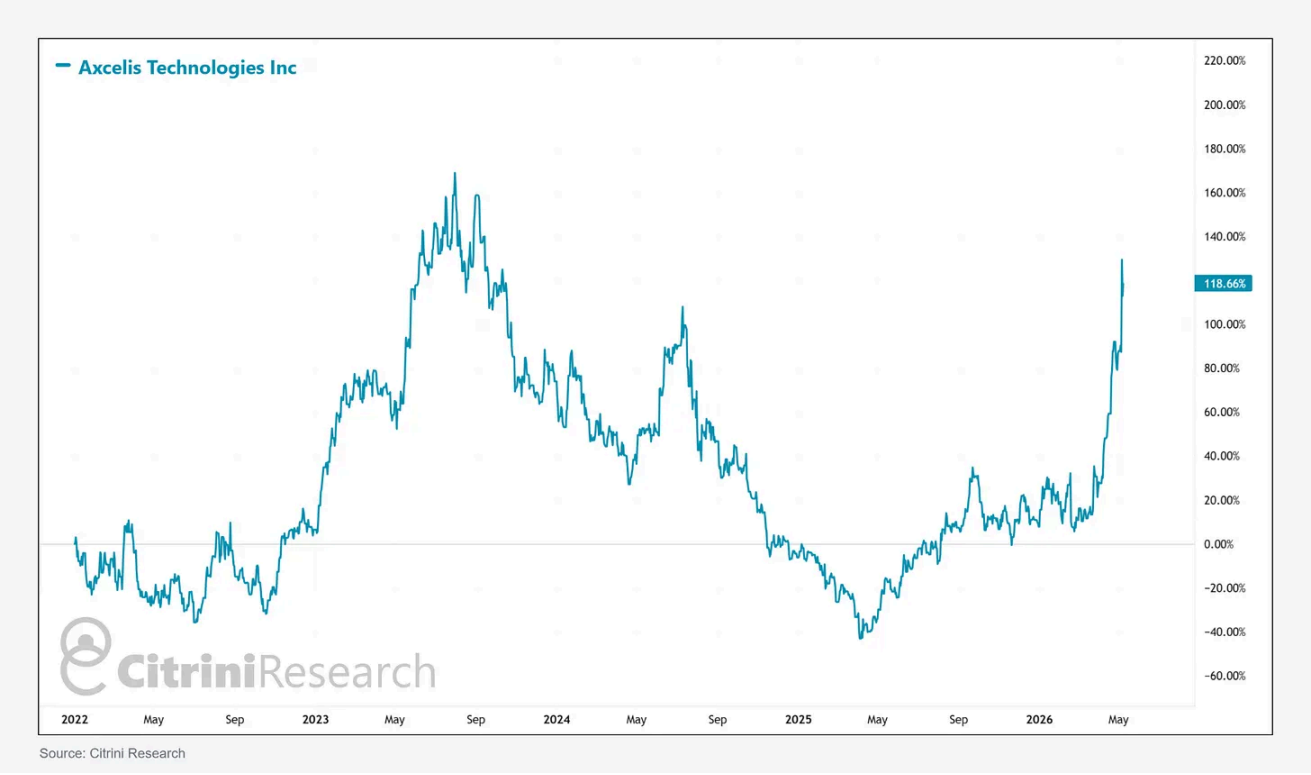
But AI infrastructure shows up as a meaningful demand layer for these names this year, with the first Rubin platform deployments at 800V, then steepens dramatically through 2028-2030 as Rubin Ultra ramps and hyperscalers commit to 800V DC for greenfield buildouts. By 2030, AI infrastructure is estimated to reach half of total SiC demand. The wafers, importantly, are the same wafers. **STMicroelectronics** (**STM US**), **ONSemi** (**ON US**), **Infineon** (**IFX GR**), etc., went from having an EV problem to

suddenly having free capacity at the exact time it was needed.

但人工智能基础设施在今年已成为这些公司的一个重要需求层，首批 Rubin 平台部署在 800V，随后随着 Rubin Ultra 的量产以及超大规模运营商在新建项目中致力于 800V 直流电，需求在 2028-2030 年间急剧增长。据估计，到 2030 年，人工智能基础设施将占碳化硅（SiC）总需求的半数。重要的是，所使用的晶圆是完全相同的。意法半导体（STM US）、安森美（ON US）、英飞凌（IFX GR）等公司，从面临电动汽车需求不足的问题，突然间在最需要的时刻拥有了闲置产能。

What makes the trade asymmetric is that the SiC names are still being valued on the EV cohort. Take a look at **Axcelis Technologies** (**ACLS US**).

这种交易之所以具有不对称性，是因为碳化硅（SiC）相关股票仍被放在电动汽车（EV）板块中进行估值。看看 Axcelis Technologies (ACLS US) 就知道了。



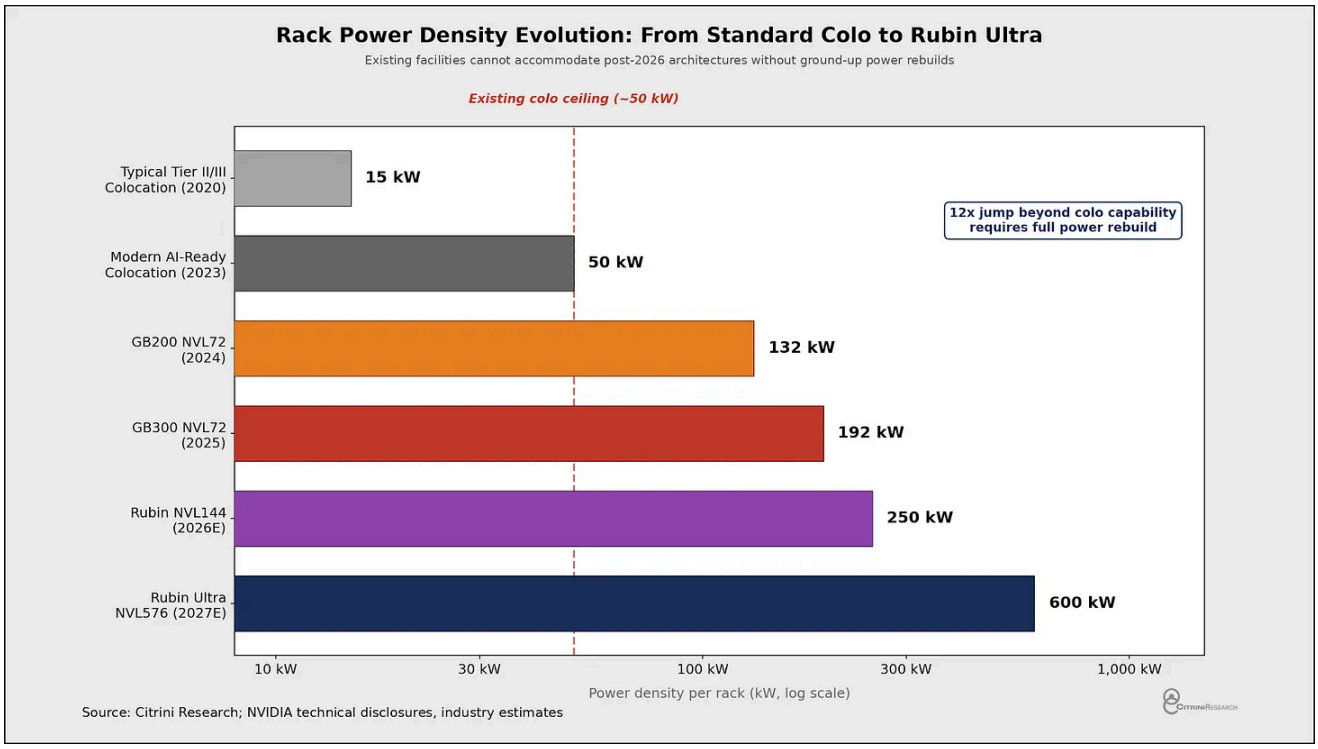
Source: Citrini Research

The next twelve months of Nvidia design wins, hyperscaler architecture commitments, and supplier earnings commentary will reveal the second cohort to the market.

未来十二个月内，Nvidia 的设计中标情况、超大规模云服务商的架构承诺以及供应商的收益评论，将向市场揭示第二批受益者。

Most colocation facilities operating today were designed for an era when 10 to 30 kW per rack was a stretch goal. The chart below puts that legacy capacity next to where Nvidia's roadmap is going. Read the x-axis carefully: it's a log scale, because the bars wouldn't fit on a linear one.

目前运营的大多数托管设施都是为每个机架 10 到 30 kW 还是挑战目标的时代而设计的。下面的图表将这些传统容量与 Nvidia 的路线图发展方向进行了对比。请仔细阅读 x 轴：它采用的是对数刻度，因为如果使用线性刻度，这些柱状图将无法在图中容纳。



A facility that was built to support 15kW per rack cannot be incrementally upgraded to support 600kW. The medium-voltage feed, the switchgear, the busways, the cooling infrastructure, the floor loading, the structural envelope itself in some cases are all wrong. In most cases, the only viable path forward is a full power infrastructure rebuild from the medium-voltage transformer inward, which is a different question from "is my colo provider expensive?" It's a question of whether the building can take the deployment at all. The large-scale operators – Digital Realty, Equinix, plus the private giants like Aligned and Compass – should have the capital and the greenfield

pipelines to build for the new spec.

一个最初为支持每机架 15kW 而设计的设施，无法通过增量升级来支持 600kW。其中压供电、开关柜、母线槽、冷却基础设施、地面承重，甚至在某些情况下连建筑结构外壳本身都是不匹配的。在大多数情况下，唯一可行的路径是从中压变压器向内进行全面的电力基础设施重建，这已经不是“我的托管服务商是否昂贵”的问题，而是该建筑是否根本无法承载这种部署的问题。大型运营商——如 Digital Realty、Equinix，以及 Aligned 和 Compass 等私有巨头——应该拥有足够的资金和绿地项目储备，来按照新规范进行建设。

Upstream: The Solid-State Transformer Inflection

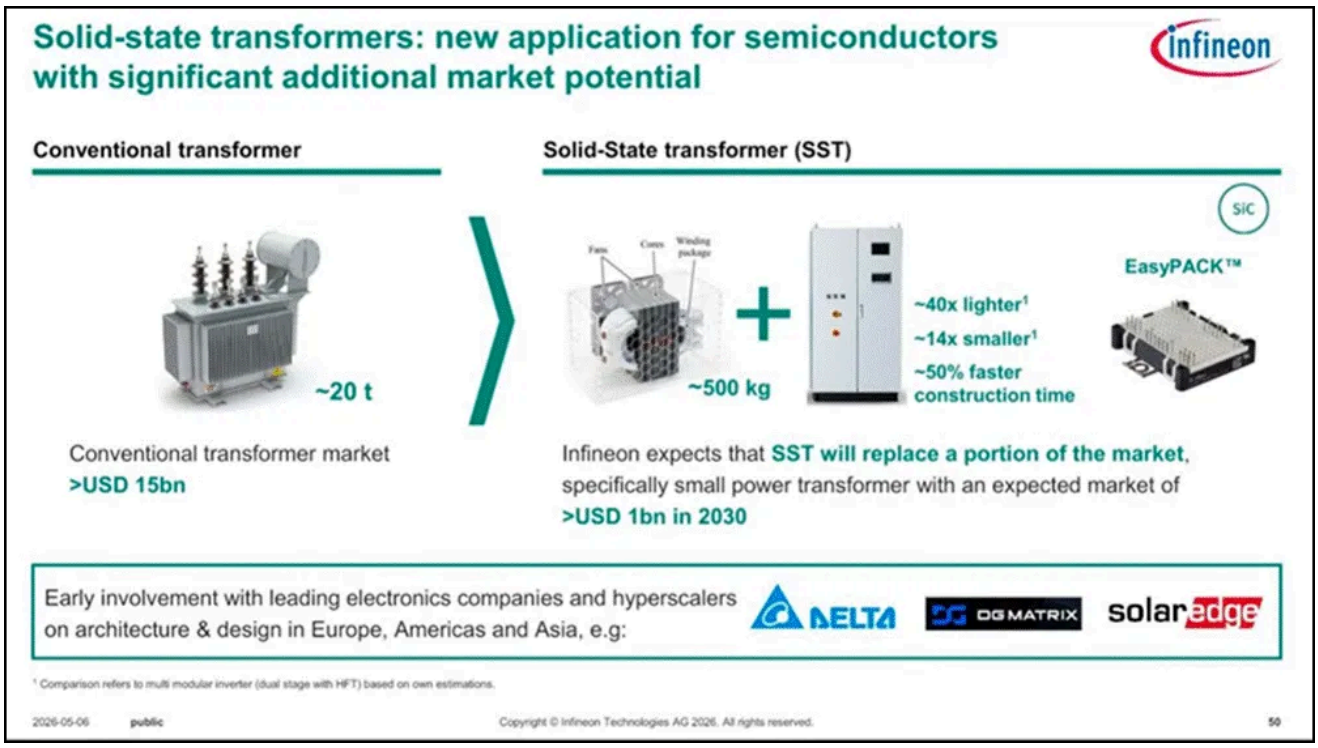
上游：固态变压器的拐点

If 800V DC distribution starts at the boundary of the data hall, something has to convert utility-grade medium-voltage AC into that 800V DC stream. Today that something is an iron-core transformer plus rectifier, a piece of equipment whose underlying technology has not meaningfully changed in a century, whose lead time is 18 to 30 months, and whose weight is measured in tens of tons.

如果 800V 直流配电从数据大厅的边界开始，就必须有设备将公用事业级的中压交流电转换为 800V 直流流。如今，这种设备是铁芯变压器加整流器，其底层技术在一个世纪内没有发生实质性变化，交付周期长达 18 到 30 个月，重量以数十吨计。

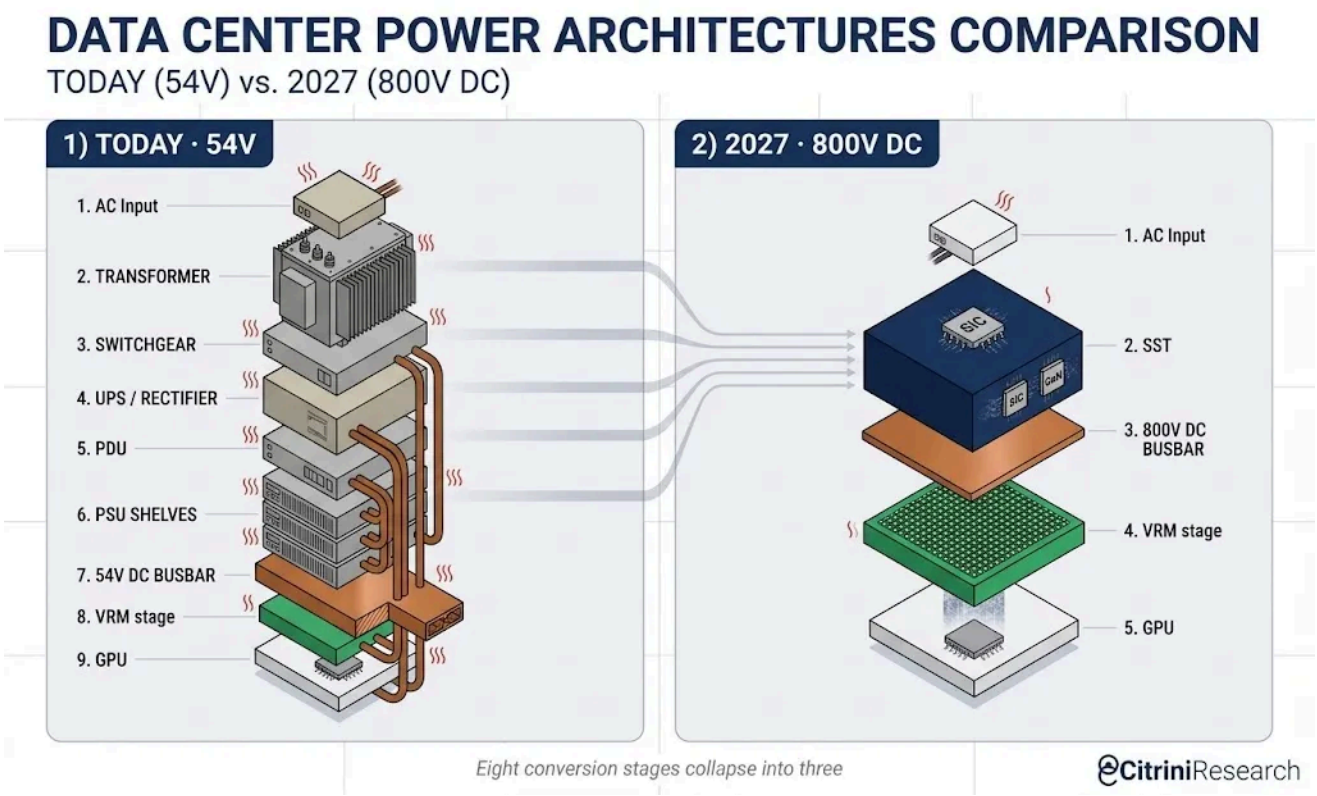
The data center industry has run out of patience for iron-core transformers. Lead times have gone from “long” to “limiting.” Major hyperscaler builds are now bottlenecked on transformer delivery rather than land, power, or chips. **The replacement technology has been quietly maturing in parallel: Solid-state transformers** (**SSTs**) are built on the same wide-bandgap semiconductor foundation as the rest of the 800V architecture.

数据中心行业对铁芯变压器的耐心已经耗尽。交付周期已从“漫长”演变为“限制性因素”。大型超大规模数据中心的建设瓶颈现在在于变压器的交付，而非土地、电力或芯片。与此同时，替代技术一直在悄然成熟：固态变压器（SST）与 800V 架构的其他部分一样，都建立在相同的宽禁带半导体基础之上。



An SST is a power electronics device, not a passive electromagnetic one. It uses high-frequency SiC switching to convert medium-voltage AC directly to 800V DC in a single integrated unit, eliminating the legacy transformer plus AC-DC rectifier pair entirely. The advantages compound across every relevant dimension: lead time (months rather than years), weight (a fraction of iron-core), physical footprint (similar reduction), functionality (built-in regulation, fault protection, bidirectional power flow), and electrical alignment with what 800V data centers actually want (native DC output).

固态变压器（SST）是一种电力电子设备，而非被动式电磁设备。它利用高频碳化硅（SiC）开关将中压交流电直接转换为 800V 直流电，集成于单一单元内，从而彻底取代了传统的“变压器加交流-直流整流器”组合。其优势体现在各个相关维度：交付周期（数月而非数年）、重量（仅为铁芯变压器的一小部分）、物理占地面积（同比例缩小）、功能性（内置调节、故障保护、双向功率流），以及与 800V 数据中心实际需求（原生直流输出）的电气匹配。



The SST opportunity falls mainly within the public companies that we've been discussing as part of the bipartisan, decades long fiscal push to modernize the grid – Eaton, ABB, Siemens etc.

SST 的机会主要集中在我们一直在讨论的那些上市公司中，这也是两党长期以来为实现电网现代化而进行的数十年财政推动的一部分——如 Eaton、ABB、Siemens 等。

Three points matter: **有三点至关重要：**

1. Traditional transformer incumbents are positioning aggressively. ABB, Eaton, Schneider, Siemens, Hitachi Energy, and GE Vernova all have publicly announced SST programs or have made acquisitions. These names are racing to get production-ready SST products to market for the 2027 deployment window and none of them have a winner-take-all position yet.

1. 传统的变压器巨头正在积极布局。ABB、伊顿 (Eaton)、施耐德 (Schneider)、西门子 (Siemens)、日立能源 (Hitachi Energy) 和通用电气 (GE Vernova) 都已公开宣布了固态变压器 (SST) 计划或进行了相关收购。这些企业正竞相在 2027 年部署窗口期前将具备生产条件的 SST 产品推向市场，目前尚无任何一家占据绝对的垄断地位。

2. Pure plays still live in the venture world. In February 2026, Heron Power (founded by Drew Baglino, ex-Tesla SVP of Powertrain and Energy) closed a $140 million Series B led by a16z's American Dynamism Fund and Breakthrough Energy Ventures, with plans to build a 40GW manufacturing facility. They claim 50GW of customer orders are already lined up. DG Matrix raised a $60 million Series A in the same window with ABB and Mitsubishi Heavy Industries participating. Eaton acquired Resilient Power last year in a deal worth up to $150 million. Unfortunately for us, there haven't been any pure-play SST IPOs.

2. 纯粹的 SST 标的仍活跃在风投领域。2026 年 2 月，由特斯拉前动力总成和能源业务高级副总裁 Drew Baglino 创立的 Heron Power 完成了由 a16z 旗下 American Dynamism 基金和 Breakthrough Energy Ventures 领投的 1.4 亿美元 B 轮融资，并计划建设一座 40GW 的制造工厂。他们声称已排产 50GW 的客户订单。DG Matrix 在同一窗口期筹集了 6000 万美元的 A 轮融资，ABB 和三菱重工 (Mitsubishi Heavy Industries) 参投。伊顿去年以高达 1.5 亿美元的价格收购了 Resilient Power。遗憾的是，目前还没有任何纯粹的 SST 公司进行 IPO。

3. SSTs widen the addressable market beyond AI capex. The same equipment that converts utility AC to 800V DC inside a hyperscale build is the equipment utilities need to modernize substations, integrate distributed renewable generation, and manage bidirectional EV charging loads. The customer base is materially bigger than the data center alone. That makes the SST piece of the basket less levered to AI capex digestion than the rest of the trade, which matters for downside protection.

3. 固态变压器（SST）将潜在市场扩展到了人工智能资本支出之外。在超大规模数据中心内部将公用事业交流电转换为 800V 直流电的同类设备，正是公用事业部门实现变电站现代化、整合分布式可再生能源发电以及管理双向电动汽车充电负载所需的设备。其客户群实质上比单纯的数据中心要大得多。这使得该组合中的 SST 部分对人工智能资本支出消化周期的依赖程度低于交易的其他部分，这对于下行保护至关重要。

Downstream: VRMs and the Thousand-Amp Problem

下游环节：电压调节模块（VRM）与千安培难题

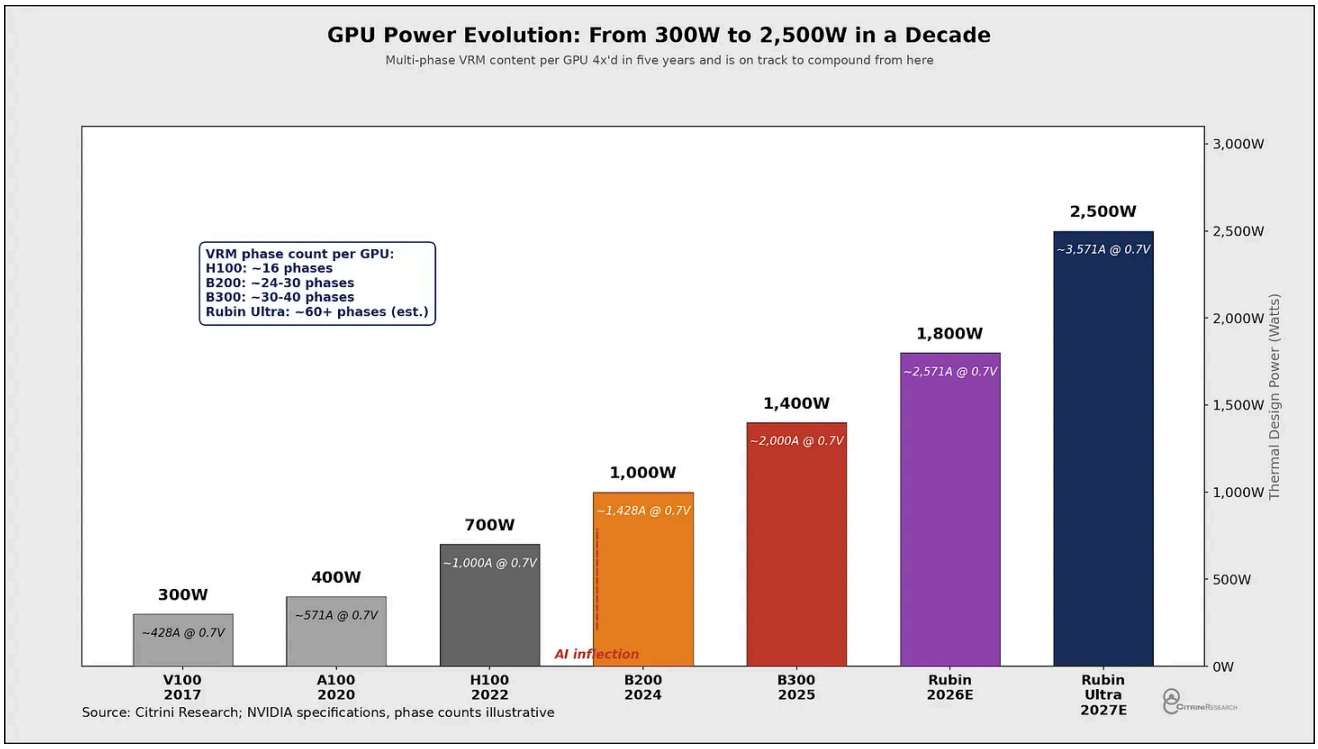
Voltage regulator modules (VRMs) are the downstream story. They sit on the motherboard immediately adjacent to the GPU, taking the intermediate bus voltage (50V, 12V, or 6V depending on architecture) and stepping it down to the sub-1V core

voltage that the silicon actually uses.

电压调节模块（VRM）是下游环节的核心。它们位于主板上紧邻 GPU 的位置，负责将中间总线电压（根据架构不同，分别为 50V、12V 或 6V）降压至芯片实际使用的 1V 以下的核心电压。

The math of GPU power delivery has gone genuinely strange. A Rubin Ultra GPU is projected to draw upward of 2,500W. At a 0.7V core voltage, the VRM has to push roughly **3,500 amps** into the silicon — not across transmission lines or factory busbars, but across a power plane the size of a paperback book.

GPU 供电的数学逻辑已经变得异常离谱。据预测，一颗 Rubin Ultra GPU 的功耗将超过 2,500W。在 0.7V 的核心电压下，VRM 必须向芯片输送约 3,500 安培的电流——这并非通过输电线或工厂母线传输，而是要通过一块仅有平装书大小的电源平面。



GPU TDP evolution and the implied current draw at the silicon. The thousand-amp problem is a five-year-old problem; the three-thousand-amp problem starts in 2027.

GPU TDP 的演进以及硅片端隐含的电流消耗。千安培问题是五年前的老问题；而三千安培问题将于 2027 年开始显现。

There is no single component that can handle that current density. The architectural response is multi-phase power delivery: dozens of parallel VRM channels each handling a fraction of the total load, synchronized in time to deliver clean current to the silicon. Phase counts have been exploding generation over generation, having quadrupled over the five years from H100 to Rubin Ultra, and each phase requires a controller IC, a high-side MOSFET, a low-side MOSFET (often integrated as DrMOS), inductors, and passive components.

没有任何单一组件能够处理如此高的电流密度。架构上的应对方案是多相供电：数十个并行的 VRM 通道各自承担总负载的一小部分，并在时间上保持同步，从而为芯片提供纯净的电流。相数在每一代产品中都呈爆炸式增长，从 H100 到 Rubin Ultra 的五年间翻了两番，而每一相都需要一个控制器 IC、一个高边 MOSFET、一个低边 MOSFET（通常集成位 DrMOS）、电感器以及被动元件。

Multiply phase count growth by the GPU volume curve and the VRM content per AI server is growing faster than almost any other category of semiconductor demand. Nvidia shipped roughly 4 million H100-class units in 2024 and is on track to ship 8-10 million Blackwell-class units in 2026. Each generation roughly doubles the units shipped and roughly doubles the VRM content per unit.

将相数的增长乘以 GPU 的出货量曲线，AI 服务器中的 VRM 价值量增长速度几乎超过了半导体需求中的任何其他类别。Nvidia 在 2024 年出货了约 400 万颗 H100 级单元，并有望在 2026 年出货 800 万至 1000 万颗 Blackwell 级单元。每一代产品的出货量大约翻倍，且单台设备的 VRM 价值量也大约翻倍。

VRMs are also where the 800V transition matters most for downstream supplier share. In the legacy 54V architecture, VRMs perform a single step-down from 54V to ~1V. In the new 800V architecture, the intermediate bus voltage drops further (to 6V-12V depending on topology), which makes the per-phase step ratio slightly easier, but the total power being delivered is higher, so phase counts keep climbing. Net effect: more content per GPU, with share captured by whoever has the best efficiency, fastest

transient response, and smallest physical footprint.

VRM 也是 800V 转型对下游供应商份额影响最大的领域。在传统的 54V 架构中，VRM 执行从 54V 到约 1V 的单级降压。在新的 800V 架构中，中间总线电压进一步下降（根据拓扑结构不同降至 6V-12V），这使得每相的压降比略微容易实现，但由于交付的总功率更高，相数仍在持续攀升。最终结果是：每颗 GPU 的价值量更高，而市场份额将被那些拥有最高效率、最快瞬态响应和最小物理占用空间的厂商夺取。

Three names dominate the AI VRM market today. **Monolithic Power Systems** (**MPWR US**) is the incumbent supplier to Nvidia's reference designs and has been the cleanest pure-play AI VRM expression for two years running. **Vicor** (**VICR US**) has been pushing factorized power architecture and vertical power delivery as alternative topologies, with mixed historical traction but growing relevance as power densities cross thresholds where conventional topologies hit physical limits. **Renesas** (**6723 JP**), via the Transphorm acquisition, is moving into GaN-based VRM topologies that offer better transient response at high phase counts. The discrete power semis (ROHM, AOSL) capture content on the MOSFET side, but have higher risk still from EVs.

如今，三家公司主导着 AI VRM（电压调节模块）市场。Monolithic Power Systems (MPWR US) 是 Nvidia 参考设计的现有供应商，连续两年以来一直是表现最纯粹的 AI VRM 标的。Vicor (VICR US) 一直在推行分级电源架构和垂直供电作为替代拓扑结构，虽然历史表现参差不齐，但随着功率密度跨越传统拓扑结构触及物理极限的阈值，其相关性正在日益增强。瑞萨电子 (Renesas, 6723 JP) 通过收购 Transphorm，正转向基于 GaN（氮化镓）的 VRM 拓扑结构，这种结构在高相数下能提供更好的瞬态响应。分立功率半导体厂商（ROHM, AOSL）在 MOSFET 端获取价值，但仍面临来自电动汽车（EV）领域的较高风险。

The Basket 一篮子股票

In order to capture this admittedly wide universe, we've split the opportunities into four tiers of the value chain, mapping closely to Nvidia's own 29-member 800VDC supplier alliance (announced at COMPUTEX 2025 and expanded through the October 2025 OCP Global Summit).

为了涵盖这个公认的广泛领域，我们将投资机会划分为价值链的四个层级，这与 Nvidia 自身的 29 成员 800VDC 供应商联盟（于 2025 年 COMPUTEX 大会上宣布，并在 2025 年 10 月的 OCP 全球峰会上进一步扩大）紧密对应。

Our conviction in the asymmetry descends by tier - most in Tier 3 and 4 have already re-rated and never had much of the EV overhang to begin with. Most of the names in Tier 1 still have yet to come close to their peak valuations from EV-mania.

Each tier captures a different slice of the share shift, and each carries a different conviction.

每一层级都捕捉到了份额转移的不同部分，且每一层级都承载着不同的信心指数。

- **Tier 1 (50% weight): Wide-bandgap silicon.** Pure-play SiC and GaN producers, plus the capital equipment vendors who sell into them.

  第一梯队（权重 50%）：宽禁带硅。纯 SiC（碳化硅）和 GaN（氮化镓）生产商，以及向其供应产品的资本设备供应商。

- **Tier 2 (25% weight): Broad-line power semi & analog.** Diversified semiconductor companies with SiC, GaN, and silicon power exposure, plus the VRM specialists driving the thousand-amp problem on motherboards.

  第二梯队（权重 25%）：通用功率半导体与模拟芯片。业务涵盖 SiC、GaN 和硅基功率器件的多元化半导体公司，以及致力于解决主板千安培电流问题的 VRM（电压调节模组）专家。

- **Tier 3 (15% weight): Power system components.** The PSU, module, busbar, and interconnect ecosystem.

  第三梯队（权重 15%）：电源系统组件。包括 PSU（电源供应器）、模块、母线排（busbar）和互连生态系统。

- **Tier 4 (10% weight): Data center power systems and medium-voltage infrastructure.** Sidecar power shelves, medium-voltage switchgear, transformers, and the SST programs that will replace iron-core transformers at the data center boundary. This is the most consensus-discovered slice of the trade, but the content shift is real and the multiples can still expand.

  第四层级（权重 10%）：数据中心电源系统和中压基础设施。包括侧置电源机架、中压开关柜、变压器，以及将在数据中心边界取代铁芯变压器的 SST（固态变压器）项目。这是该交易中最被共识发掘的部分，但内容增量是真实存在的，且估值倍数仍有扩张空间。

ANALOG & POWER SEMIS

AI Rack Architecture · Wide-Bandgap, VRM & Power Infrastructure Basket

WIDE-BANDGAP SEMICONDUCTORS

SiC / GaN · Device Makers, Foundries, Modules & Equipment

TIER 1 (50%)

Infineon Technologies — IFX GR
SiC/GaN power leader across devices, modules and drivers

STMicroelectronics — STM US
Integrated SiC supplier bridging EV, industrial and datacenter power

onsemi — ON US
EliteSiC portfolio for high-efficiency AI power conversion

ROHM — 6963 JP
SiC MOSFET and module specialist with deep power pedigree

Navitas Semiconductor — NVTS US
GaNFast and GeneSiC supplier selected for 800V HVDC

Power Integrations — POWI US
High-voltage GaN ICs for compact AC/DC conversion

Innoscience — 2577 HK
Scaled GaN-on-silicon platform moving into higher-power applications

Wolfspeed — WOLF US
SiC substrates and devices, pure-play but financially stressed exposure

Fuji Electric — 6504 JP
All-SiC high-power modules spanning 1200V to 3300V classes

X-FAB Silicon Foundries — XFAB FP
Specialty foundry for SiC, GaN and high-voltage power devices

Aixtron — AIXA GR
Epitaxy tools enabling GaN and SiC device capacity

Axcelis Technologies — ACLS US
Ion implantation tools for SiC power-device manufacturing

POWER MANAGEMENT, GATE DRIVERS & PROTECTION

Voltage Regulation · Gate Drivers · Mixed-Signal Power · Current Protection

TIER 2 (25%)

Monolithic Power Systems — MPWR US
High-density power modules and regulators near the GPU

Analog Devices — ADI US
Precision power, sensing and control for rack conversion

Texas Instruments — TXN US
Analog, isolation and gate-drive backbone for power stages

Alpha & Omega Semiconductor — AOSL US
Discrete MOSFETs and power ICs for high-current conversion

Vicor — VICR US
Modular power conversion specialist for AI rack architectures

Renesas Electronics — 6723 JP
Controllers, drivers and mixed-signal power management

Mersen — MRN FP
Fuses, busbars and thermal materials for high-current DC

Microchip Technology — MCHP US
mSiC devices, modules and gate drivers from 700V to 3.3kV

BUSBARS, CONNECTORS & POWER MODULES

Rack-Level Copper Architecture · Interconnect · Assembly

TIER 3 (15%)

Delta Electronics — 2308 TT
Power supplies, busbars and thermal systems for AI racks

BizLink — 3665 TT
Power cords, harnesses and cable assemblies for infrastructure

Megmeet — 002851 CH
Power supplies and conversion systems for datacenter loads

Flex Ltd — FLEX US
Design-to-manufacturing partner for rack power assemblies

Amphenol — APH US
High-current connectors and interconnects for power distribution

TE Connectivity — TEL US
Connectors, relays and sensors across rack infrastructure

nVent Electric — NVT US
Electrical enclosures, cooling and protection for power gear

Methode Electronics — MEI US
Busbar and power-distribution assemblies for high-current systems

Prysmian — PRY IM
Grid-to-rack cable systems for high-current campuses

INDUSTRIAL POWER INFRASTRUCTURE

Grid Edge · Switchgear · Transformers · Datacenter Electrification

TIER 4 (10%)





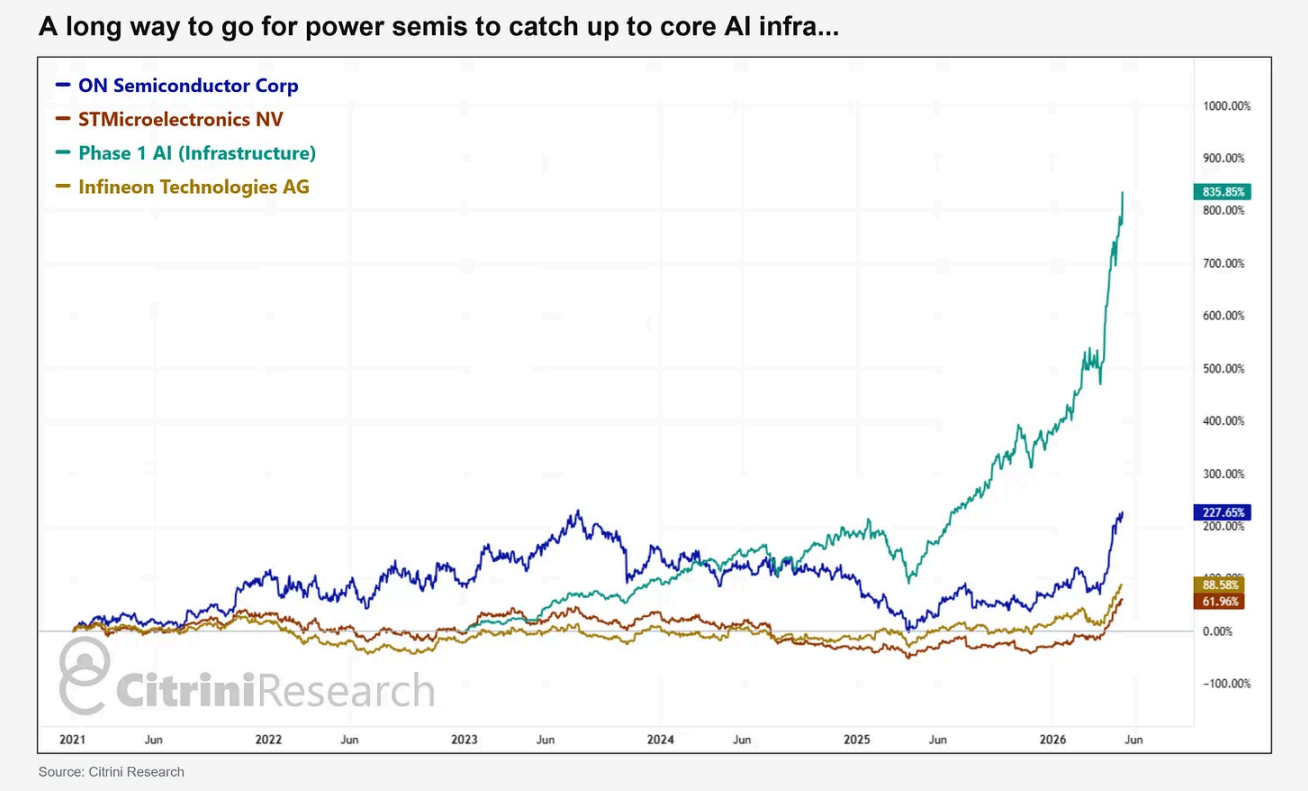
Source: Citrini Research

You can find the basket on Citrindex here.

您可以在 Citrindex 上找到该一篮子指数，链接请点击此处。

And before you scoff at buying a chart that's going vertical (although if that's your default, you're probably already having a pretty bad time), take a look at some of the most consequential names for 800V when zoomed out versus our 125-name Core AI Data Center Infrastructure basket.

在你嘲笑购买这种呈垂直上升趋势的图表之前（尽管如果这是你的默认思维方式，你现在的处境可能已经相当糟糕了），不妨将 800V 领域中一些最具影响力的公司与我们包含 125 个成分股的核心 AI 数据中心基础设施篮子进行一下远景对比。



The market might be coming around to the realization of the supply chain inheritance, but it's far from having repriced these as "AI names". Most AI infrastructure trades require the AI buildout to keep going, full stop. The supply chain inheritance trade has a second leg of demand under it (EVs, robotics, grid buildout, and industrial power) that makes it structurally insulated in a way that pure-play AI infrastructure trades aren't. We are *just now* beginning to see the market reprice them as if EVs aren't the only thing the wafers are doing.

市场可能正在逐渐意识到供应链传承的价值，但远未将这些标的作为"AI 概念股"重新定价。绝大多数 AI 基础设施交易都要求 AI 建设必须持续进行，仅此而已。而供应链传承交易下方拥有第二层需求支撑（电动汽车、机器人、电网建设和工业电力），这使其在结构上具备了纯 AI 基础设施交易所不具备的防御性。我们现在才刚刚开始看到市场对其进行重新定价，意识到这些晶圆的用途绝不仅仅局限于电动汽车。

This memo is going to be long as-is, so I will simply refer readers interested in understanding the other kicker here (upside to the power needs for embodied AI / robotics) to our **Humanoid Robots** primer and **Robotics Update** article. That second wave of demand that will hit over the coming years in addition to data centers will

make this one of the hottest semi subsectors.

这份备忘录本身篇幅已经很长了，所以对于那些有兴趣了解另一个潜在增长点（具身智能/机器人技术对电力需求的提振）的读者，我建议参考我们的《人形机器人入门指南》和《机器人行业更新》文章。除了数据中心之外，未来几年将迎来的第二波需求浪潮，将使该领域成为半导体行业中最热门的细分板块之一。

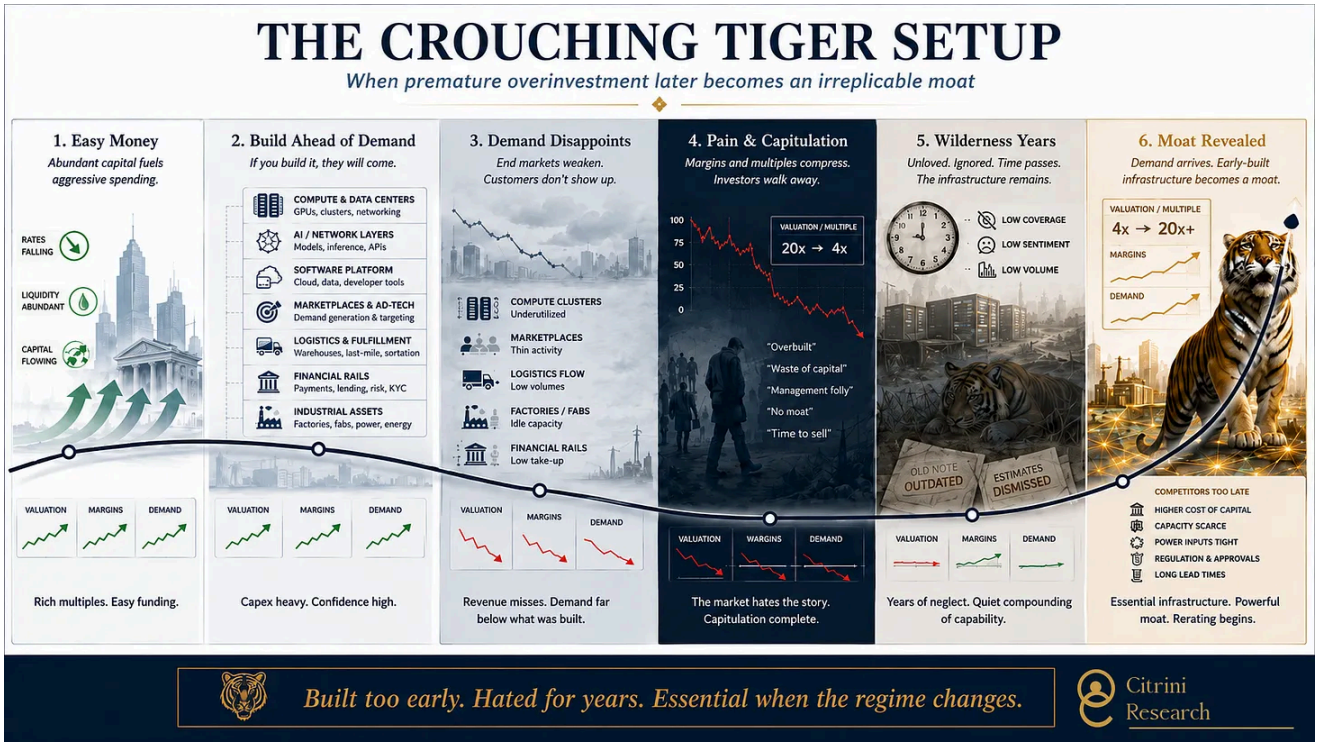
Single-Stock Highlight: Wolfspeed (WOLF US)

个股亮点：Wolfspeed (WOLF US)

There's a particular type of setup we've been very fond of for the past two years: it's the setup that we saw with our **Crouching Tigers, Hidden Dragons** in 2024-2025.

在过去的两年里，我们一直非常青睐一种特定的形态：即我们在 2024-2025 年“卧虎藏龙”系列中看到的那种形态。

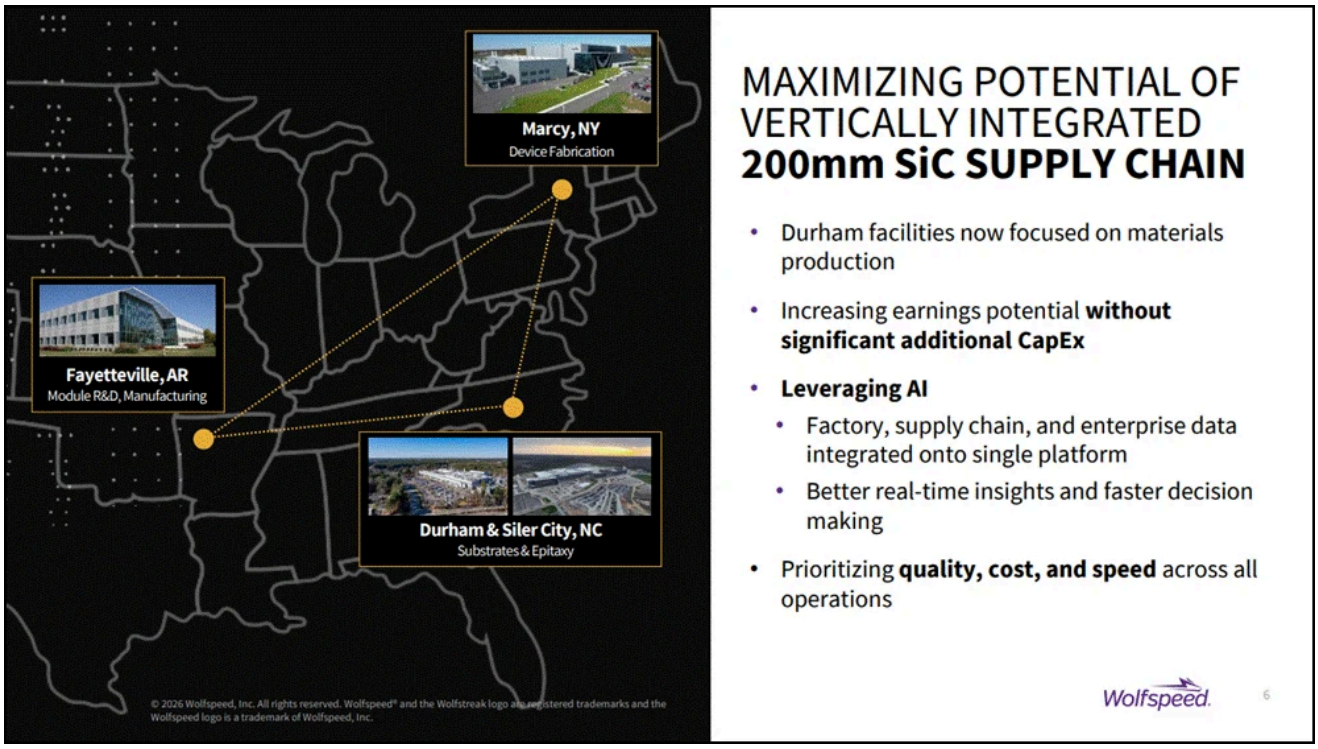
It goes like this: 具体如下：



This was our idea behind names like **Carvana** (**CVNA US**), **Sea Limited** (**SE US**), and **AppLovin** (**APP US**) heading into 2025 and now it's spilled over into the semis space. Wolfspeed is the platonic ideal – not only did they spend aggressively to ramp capacity

into demand that failed to materialize, they did it so hard that it bankrupted their company.

这就是我们在迈向 2025 年时，对 Carvana (CVNA US)、Sea Limited (SE US) 和 AppLovin (APP US) 等公司的核心思路，而现在这种逻辑已经蔓延到了半导体领域。Wolfspeed 就是一个典型的反面教材——他们不仅为了应对未能兑现的需求而激进地扩张产能，而且用力过猛，直接导致了公司的破产。



Wolfspeed spent $6.5 billion building out its global SiC capacity, anchored by the world's first 200mm SiC fab at Mohawk Valley and the John Palmour Manufacturing Center in Siler City, NC. Their bet was that EV adoption would absorb the capacity on schedule.

Wolfspeed 斥资 65 亿美元建设其全球碳化硅（SiC）产能，其核心是位于莫霍克谷（Mohawk Valley）的全球首座 200mm 碳化硅晶圆厂，以及位于北卡罗来纳州赛勒市（Siler City）的 John Palmour 制造中心。他们的赌注是电动汽车（EV）的普及将按计划消化这些产能。

EV adoption… *didn't absorb the capacity on schedule*.

电动汽车的普及……并未按计划消化掉产能。

The stock, originally CREE, went from $140 to single digits and filed Chapter 11 in June 2025. Bondholders took control in a debt-for-equity swap that wiped out the old holders. Sorry, previous WOLF equity holders. Now, however...

该股票原代码为 CREE，股价从 140 美元跌至个位数，并于 2025 年 6 月申请了第 11 章破产保护。债权人通过债转股接管了公司，导致原股东权益归零。很遗憾，之前的 WOLF 股东们。然而，现在……

The CREEdence Clearwater Revival



The setup now, on the other side of bankruptcy, is perfect. The infrastructure is coveted – Mohawk Valley is still the only 200mm SiC fab operating at commercial scale *in the entire world*. The John Palmour facility is the largest SiC wafer fab ever built. The physical capacity that took five years to build now finds itself in a place where the EV cohort doesn't really matter anymore, and the AI infrastructure cohort needs their capacity to get to the next level. And, as we all know, nothing gets in the

way of this train.

在破产重组之后，现在的局面堪称完美。其基础设施备受瞩目——Mohawk Valley 依然是全球唯一一家以商业规模运营的 200mm 碳化硅（SiC）晶圆厂。John Palmour 设施则是史上规模最大的碳化硅晶圆厂。这些耗时五年建成的物理产能，如今发现电动汽车（EV）群体的需求已不再是决定性因素，而人工智能（AI）基础设施群体正急需这些产能来迈向新台阶。而且，众所周知，没有什么能阻挡这列滚滚向前的时代列车。

The spending cycle that bankrupted them gives the restructured company an irreplaceable position in the physical supply chain that we just spent 10 pages outlining. The capex looked (and was) value destructive in 2023. By 2027, it will look like one of the most prescient infrastructure bets in the wide-bandgap industry. The market is beginning to price this as a turnaround – **we think it should be priced as an unavoidable beneficiary.**

曾令其破产的支出周期，却让重组后的公司在物理供应链中占据了不可替代的地位，而这正是我们刚刚用 10 页篇幅所概述的内容。资本支出在 2023 年看起来（且确实）是在破坏价值。但到 2027 年，它将被视为宽禁带半导体行业中最具先见之明的基础设施赌注之一。市场正开始将其定价为一场困境反转——而我们认为，它应该被定价为不可或缺的受益者。

That's our base case, right now. WOLF is a crouching tiger getting ready to reveal a dragon that deserves to not just be priced based on what their fab's replacement value theoretically is, but reflect the fact that it's not going to be replaced. By anyone. They're the only game in town.

这就是我们目前的基本假设。WOLF 是一只卧虎，正准备展现出其真龙之姿；它的定价不应仅仅基于其晶圆厂理论上的重置价值，而应反映出一个事实：它将是无可替代的。没有任何人能取代它。他们是这个领域唯一的玩家。

But wait, there's more. **别急，还没完。**

Beyond getting paid as AI infrastructure absorbs that 200mm capacity, there's the 300mm kicker where either WOLF will repeat the exact same mistakes (which would

be kind of funny) or cement its lead for decades to come.

除了随着 AI 基础设施吸收 200mm 产能而获得回报外，还有一个 300mm 的潜在利好：要么 WOLF 会重蹈覆辙（那确实有点滑稽），要么它将巩固未来几十年的领先地位。

First, a disclaimer: there are no 300mm Silicon Carbide wafer production or fabrication facilities online anywhere in the world. Market consensus sees this as a 6-8 year roadmap with more clear risks than reward, we just saw $10 billion of capital destruction for a single 200mm buildout, so why the optimism here? ***Why now?***

首先，郑重声明：目前全球范围内还没有任何投入运营的 300mm 碳化硅（SiC）晶圆生产或制造设施。市场共识认为这是一个为期 6 到 8 年的路线图，且风险明显大于回报——我们刚刚目睹了单个 200mm 扩产项目就造成了 100 亿美元的资本毁灭，那么为什么现在要保持乐观？又为什么是现在？

Because the same property that made SiC useful at the edges of a data center (absurd thermal conductivity) makes it a candidate to receive an invitation much closer to the silicon itself.

因为正是那种让 SiC 在数据中心边缘大显身手的特性（极高的导热率），使其有望被引入到更接近硅芯片核心的位置。

As we have said repeatedly in our coverage of the upstream to downstream AI material bottlenecks - **The closer a material sits to the chip, the more pricing leverage its vendor has over the customer.** At the periphery, SiC competes on price like any other power material. Inside the package, the economics flip entirely.

正如我们在对 AI 材料从上游到下游瓶颈的多次报道中所述：一种材料距离芯片越近，其供应商对客户的定价权就越大。在边缘地带，SiC 像其他任何功率材料一样进行价格竞争；但在封装内部，经济逻辑将完全反转。

The catalyst here is thermal. Modern AI packages are pushing 1,000W+ through a footprint smaller than a credit card, and current cold-plates (copper or aluminum, sitting on top of a thermal interface material on top of the die) are a tower of bottlenecks. SiC has roughly 3x the thermal conductivity of copper once you account for metallization stack penalties, and dropping a SiC cold-plate directly onto the back

of a die assembly improves the heat path dramatically.

这里的催化因素是热管理。现代 AI 封装在比信用卡还小的面积上产生超过 1,000W 的功耗，而目前的冷板（铜或铝材质，置于芯片顶部的热界面材料之上）存在重重瓶颈。一旦考虑到金属化堆栈的损耗，碳化硅（SiC）的热导率大约是铜的 3 倍，将 SiC 冷板直接置于芯片组件背面可以显著改善散热路径。

The natural extension is to machine fluid cooling channels directly into the SiC substrate itself, which is the active cooling concept that NASA was funding a private American company called Eotron to demonstrate over a decade ago. Eotron is still doing that work today. The proposed “Silicon Cold Plate” sits in direct contact with high thermal flux devices and removes waste heat through internal fluid channels, while also routing electrical and system connections through the same substrate.

自然的延伸是在 SiC 基板本身直接加工出流体冷却通道，这是美国宇航局（NASA）在十多年前资助一家名为 Eotron 的美国私营公司所论证的主动冷却概念。Eotron 至今仍在从事这项工作。提议中的“硅冷板”与高热通量器件直接接触，通过内部流体通道带走废热，同时在同一基板上布设电气和系统连接。

Take this one step further and chiplet interposers themselves get fabbed from SiC, with through-vias and communication fabrics carved directly into a substrate whose backside is doing the thermal work. We know the science is real. We know the demand will be enormous. So why hasn’t anyone built this yet?

再进一步，芯粒（chiplet）中介层本身也可以由 SiC 制造，通过通孔和通信架构直接刻蚀在背面承担散热任务的基板中。我们知道这项科学技术是真实的，也知道需求将是巨大的。那么，为什么还没有人把它造出来呢?

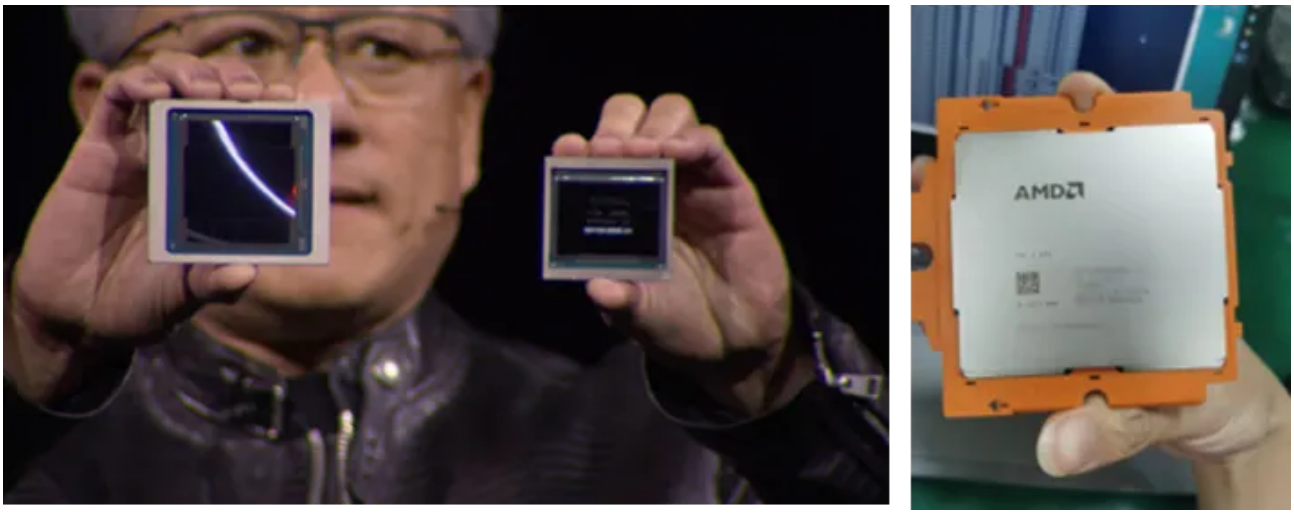
Geometry. Straight up sixth grade geometry.

几何学。纯粹就是六年级的几何学。

If you missed our discussion about the importance of advanced packaging in 26 Trades for 2026, well, sorry, you could have made some serious money if you hadn’t, but we’ll refresh quickly. Modern AI accelerators have outgrown the reticle (the hard physical cap on the size of a single die, roughly 850mm^2 for EUV). Once you exceed that limit, you have to use chiplets (a bunch of smaller chips packaged together), which means you need a substrate or interposer big enough to host multiple chiplets next to each

other. The AMD MI300A package lands around 75x72mm. Blackwell GPU packages are roughly 70x76mm. We are not done climbing.

如果你错过了我们在《2026 年的 26 笔交易》中关于先进封装重要性的讨论，那么很遗憾，如果你当时关注了，本可以赚到大钱，但我们会快速回顾一下。现代 AI 加速器已经超出了光罩限制（reticle limit，即单颗芯片尺寸的物理硬上限，对于 EUV 工艺约为 850mm^2）。一旦超过这个限制，就必须使用小芯片（chiplets，即封装在一起的一组较小芯片），这意味着你需要一个足够大的基板或中介层，以便将多个小芯片并排容纳。AMD MI300A 的封装尺寸约为 75x72mm。Blackwell GPU 的封装尺寸约为 70x76mm。而这种增长趋势尚未见顶。

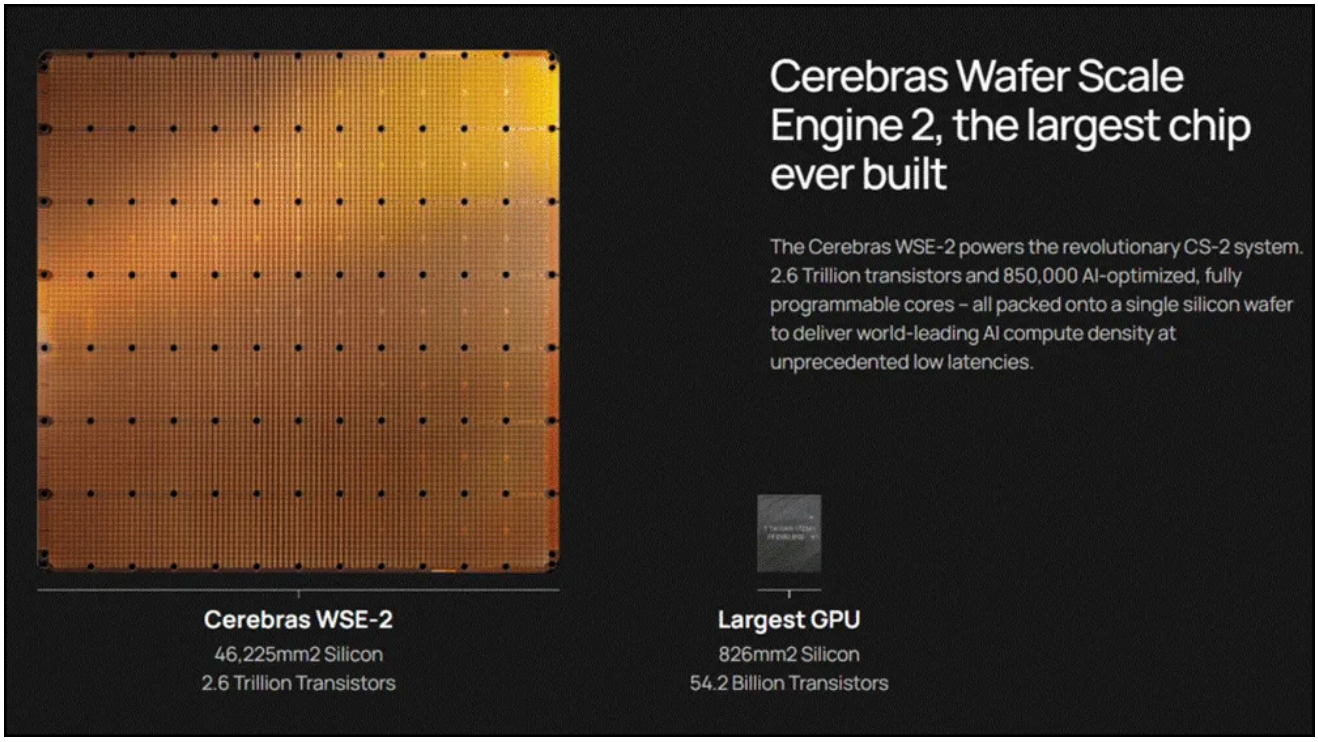


Jensen Huang holds a Grace CPU (your left), a Blackwell GPU die (center), and a fab tech holds an extra-large AMD package (your right, not to scale)

黄仁勋手持一颗 Grace CPU（左侧）、一颗 Blackwell GPU 裸片（中间），一名晶圆厂技术人员手持一个超大尺寸的 AMD 封装（右侧，非等比例）

The extreme case is Cerebras, which skipped chipleting altogether and claimed an entire wafer as one die. As if a boomer product manager showed up one day and said “what if we just made one big chip?” and the engineers said “no you can’t do…well, I mean, okay yeah technically that could work”.

极端案例是 Cerebras，它完全跳过了芯粒化（chipleting），直接将整块晶圆宣称为一颗芯片。这就像是一个婴儿潮一代的产品经理某天突然出现并问道：“如果我们只做一个超大芯片会怎样？”而工程师们回答说：“不，你不能这么……好吧，我的意思是，行吧，技术上这确实可行。”

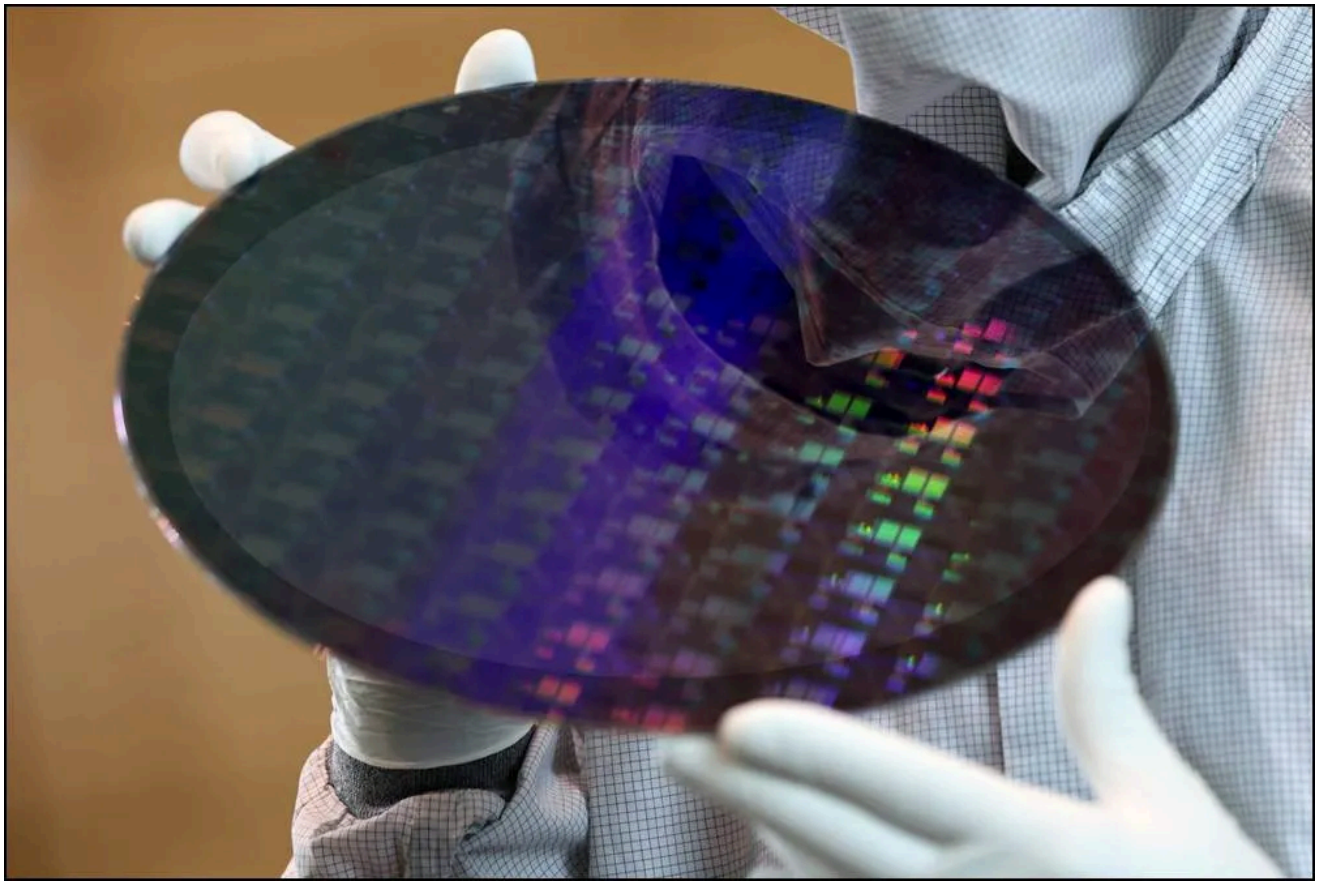


But notice anything about these chips when compared to wafers (which you'll recognize the shape of from TSMC logo fame).

但请注意，将这些芯片与晶圆（你会从台积电著名的标志中认出其形状）进行对比时，是否发现了什么。

A wafer is a round disk. The chips you want to cut out of it are squares. That's a problem, because you can't start your square at the edge of a circular wafer (you need clean straight edges so the squares can be stacked, bonded, or grown vertically into other dies without defects propagating from a curved boundary).

晶圆是一个圆盘。而你想从中切割出的芯片是正方形的。这便产生了一个问题，因为你无法从圆形晶圆的边缘开始切割正方形（你需要整齐的直边，以便正方形可以进行堆叠、键合或垂直生长成其他晶粒，而不会让缺陷从弧形边界蔓延开来）。

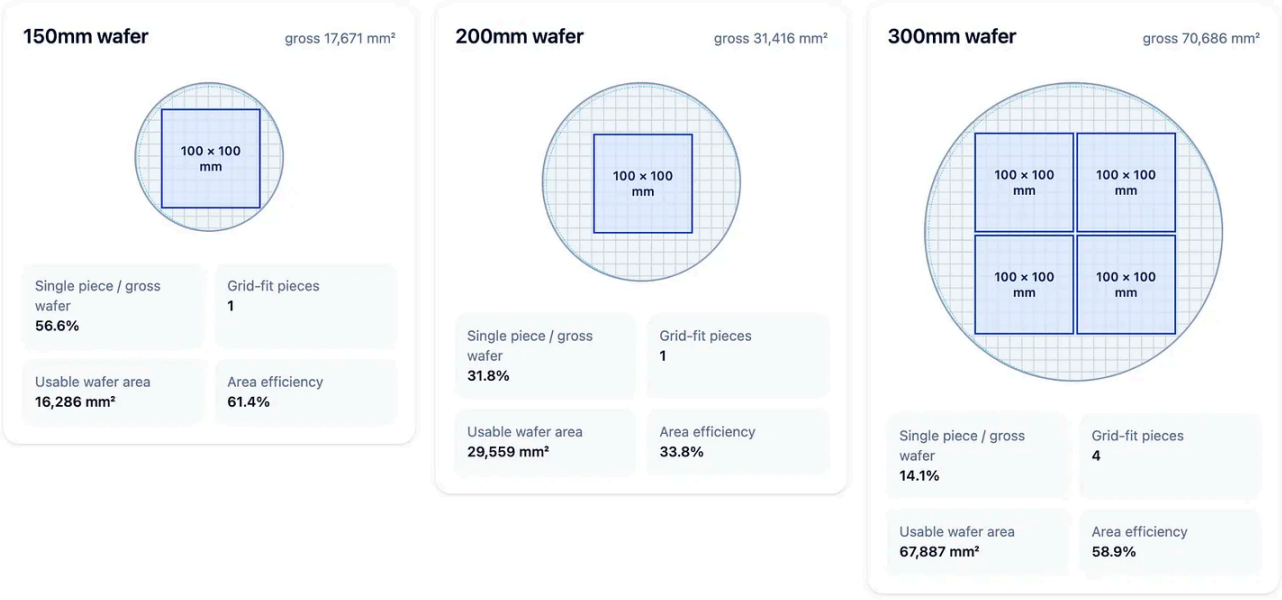


Wafers already throw off a lot of wasted material, this would take that to the extreme even before accounting for the impact a single defect could have.

晶圆本身已经会产生大量废料，而在尚未考虑单个缺陷可能带来的影响之前，这种情况就已经达到了极致。

Bump the diameter up by 50% to 300mm and suddenly you fit four of those packages on the same wafer. The math moves from “wildly uneconomic” to “viable product.”

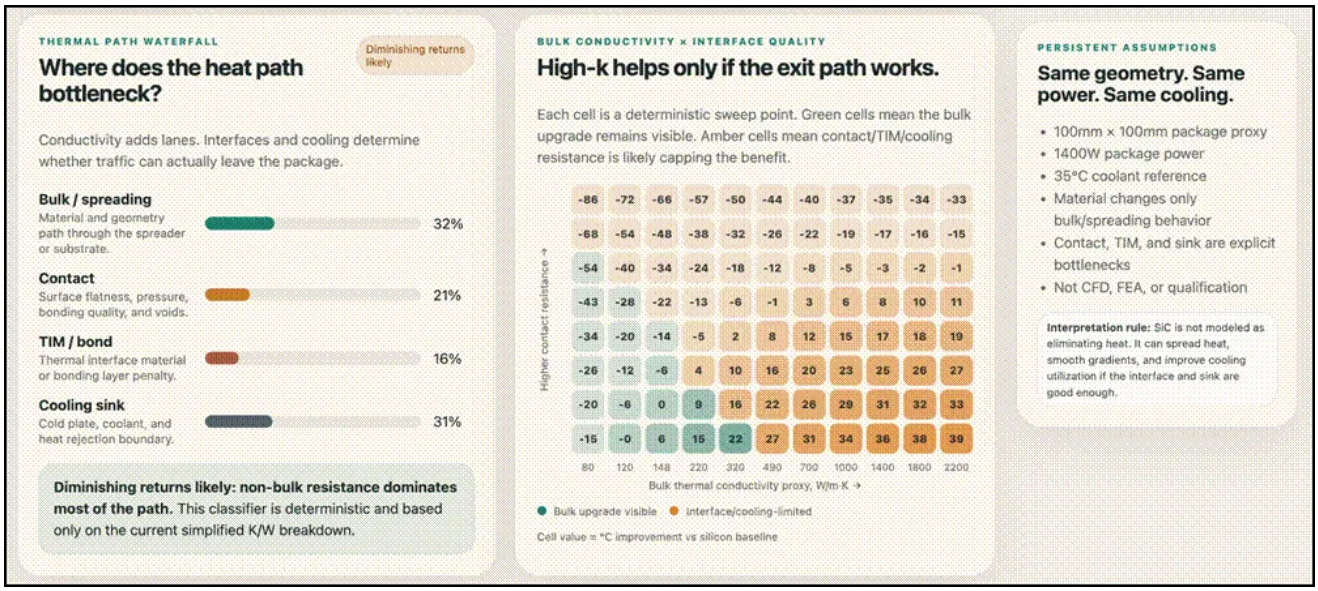
将直径增加 50% 至 300mm，突然之间你就能在同一片晶圆上容纳四个那样的封装。数学计算结果便从“极度不经济”转变为“可行产品”。



Read-through

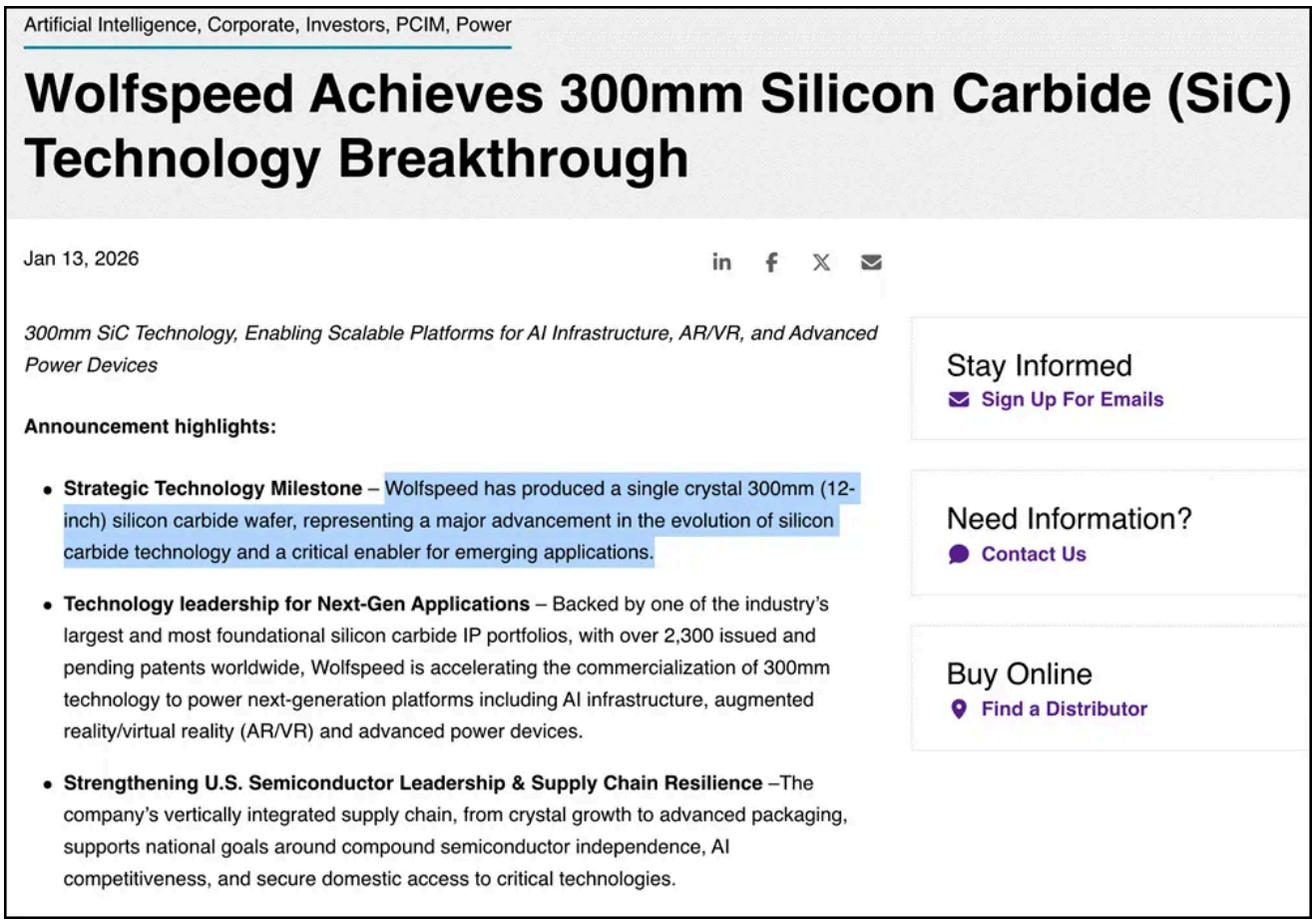
Wafer	Gross area	Gross area vs. 150mm	Grid-fit pieces	Single piece / gross area	Interpretation
150mm	17,671 mm²	1.00×	1	56.6%	Very constrained for package-scale SiC pieces.
200mm	31,416 mm²	1.78×	1	31.8%	Technically feasible but poor geometry for very large components.
300mm	70,686 mm²	4.00×	4	14.1%	Much better geometry for scarce, premium package-scale components.

Simplified model. It does not account for wafer notch, real die-placement optimization, defect maps, bow/warp, wafer thickness, process exclusion zones, metallization/test coupons, component shape variants, or yield-learning effects.



The obvious next question: is Wolfspeed working on this, and did the relevant assets survive the bankruptcy? Yes, and yes.

接下来的问题显而易见：Wolfspeed 是否正在研发这项技术，以及相关资产是否在破产中幸存？答案是肯定的，两个问题的回答都是肯定的。



Artificial Intelligence, Corporate, Investors, PCIM, Power

Wolfspeed Achieves 300mm Silicon Carbide (SiC) Technology Breakthrough

Jan 13, 2026

300mm SiC Technology, Enabling Scalable Platforms for AI Infrastructure, AR/VR, and Advanced Power Devices

Announcement highlights:

- **Strategic Technology Milestone** – Wolfspeed has produced a single crystal 300mm (12-inch) silicon carbide wafer, representing a major advancement in the evolution of silicon carbide technology and a critical enabler for emerging applications.
- **Technology leadership for Next-Gen Applications** – Backed by one of the industry's largest and most foundational silicon carbide IP portfolios, with over 2,300 issued and pending patents worldwide, Wolfspeed is accelerating the commercialization of 300mm technology to power next-generation platforms including AI infrastructure, augmented reality/virtual reality (AR/VR) and advanced power devices.
- **Strengthening U.S. Semiconductor Leadership & Supply Chain Resilience** –The company's vertically integrated supply chain, from crystal growth to advanced packaging, supports national goals around compound semiconductor independence, AI competitiveness, and secure domestic access to critical technologies.

Stay Informed
Sign Up For Emails

Need Information?
Contact Us

Buy Online
Find a Distributor

In January 2026, the post-restructuring company demonstrated production of a single 300mm SiC crystal.

2026 年 1 月，重组后的公司展示了单晶 300mm 碳化硅（SiC）晶体的生产能力。

Yes, it's only one wafer. But Wolfspeed is one of only two companies on earth that has shown this capability (Coherent is the other), and Wolfspeed is the only one of those two whose entire business identity is wide-bandgap materials. Coherent has SiC as a division. Wolfspeed ***is*** SiC.

是的，这仅仅是一片晶圆。但 Wolfspeed 是全球仅有的两家展示出这种能力的公司之一（另一家是 Coherent），而且在两家公司中，Wolfspeed 是唯一一家将整个业务身份都押注在宽禁带材料上的公司。Coherent 只是将 SiC 作为一个部门，而 Wolfspeed 本身就是 SiC。

Current equity holders are free-and-clear of the debt overhang from the previous (EV focused) build-out and are primed to ride the coattails of larger-die AI accelerators to the heart of the data center rather than just the periphery. If the story were just about EVs, 300mm SiC wafers are probably an early-2030s reality, but the massive margins

and urgent aggression of frontier AI companies could easily pull this forward to reality before the 2020s wind down.

目前的股权持有者已经摆脱了之前（专注于电动汽车）扩建所带来的债务阴影，并准备好顺着大尺寸芯片 AI 加速器的东风，从数据中心的边缘进入其核心地带。如果故事仅仅关乎电动汽车，300mm SiC 晶圆可能要到 2030 年代初期才能成为现实；但前沿 AI 公司巨大的利润空间和紧迫的进取心，很可能会在 2020 年代结束前就将其变为现实。

The cherry on top is that New York State has 'invested' billions of dollars off of the balance sheet and does not expect direct repayment, so the deck is stacked in our favor on this one. The MTA alone needs to replace 190 4MW "power traction" substations just for the NYC subway while 300+ miles of track on LIRR/MNR are being electrified as well as a healthy rolling stock cycle. Then you have city and school buses going zero emission by 2030 and the charging infrastructure required for that.

锦上添花的是，纽约州已经在资产负债表外"投资"了数十亿美元，且不期望直接还款，因此在这件事上形势对我们非常有利。单是纽约大都会运输署（MTA）就需要为纽约地铁更换 190 个 4MW 的"动力牵引"变电站，同时长岛铁路（LIRR）和北铁（MNR）还有 300 多英里的轨道正在进行电气化改造，此外还有庞大的车辆更新周期。再加上城市和学校巴士将在 2030 年前实现零排放，以及为此所需的充电基础设施。

Oh, and let us leave you with one more hook: Elon Musk will need a metric ton of these components and has a vested interest in promoting American manufacturing.

哦，对了，再给各位留一个悬念：埃隆·马斯克将需要大量的这类组件，而且他在推动美国本土制造方面有着切身的利益。

CPUs in the Agentic Era

智能体时代的 CPU

As we covered in Agentic Utilities – the next phase of AI is about models that can do things. A chatbot answers a prompt or query. An agent completes a task. It searches,

plans, retrieves, calls APIs, runs code, checks results, updates memory, coordinates with other agents, and then calls the model again. The model is still essential, but it is no longer the whole workload. The workload has transformed into a loop and that loop needs orchestration.

正如我们在《代理化实用程序》（Agentic Utilities）中所探讨的——人工智能的下一阶段是关于能够“做事”的模型。聊天机器人回答提示或查询，而代理（Agent）则完成任务。它进行搜索、规划、检索、调用 API、运行代码、检查结果、更新记忆、与其他代理协调，然后再次调用模型。模型依然至关重要，但它不再是全部的工作负载。工作负载已经转化为一个循环，而这个循环需要编排。

Conveniently, this is where the CPU comes back into the AI story.

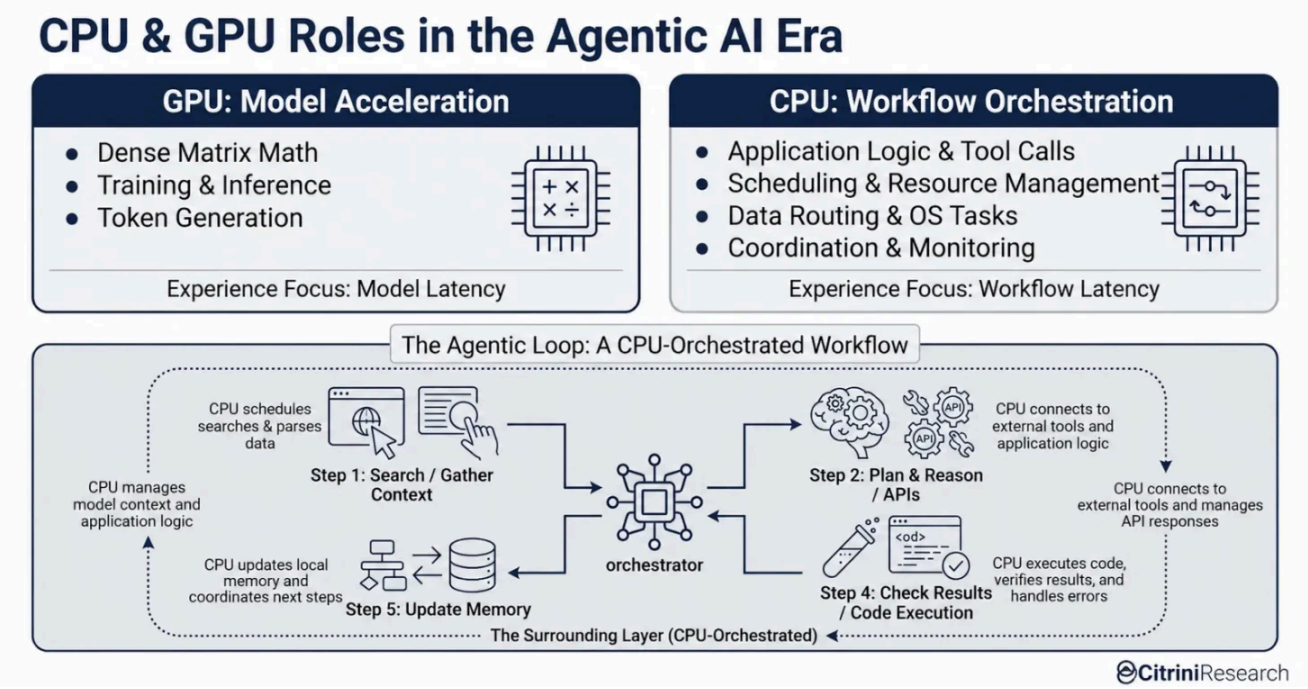
顺理成章地，这正是 CPU 重新回到 AI 叙事中的时刻。

GPUs generate tokens; CPUs keep the agents moving. Agents tend to use 10-30x more tokens than chatbot interfaces. In the chatbot era, the user mostly experienced model latency. In the agentic era, the user now experiences workflow latency. The workload now involves an agent gathering context, calling tools, checking outputs, completing actions, and returning something useful to the user.

GPU 生成 Token；CPU 则维持代理的运转。代理使用的 Token 数量往往比聊天机器人界面多出 10 到 30 倍。在聊天机器人时代，用户主要感受到的是模型延迟。而在代理时代，用户现在感受到的是工作流延迟。现在的工作负载涉及代理收集上下文、调用工具、检查输出、完成操作，并向用户返回有用的结果。

That surrounding layer is where the CPU lives.

那一外层就是 CPU 所在的区域。



As we laid out in **Interconnects 101**, a CPU isn't equipped to do the dense matrix math that makes GPUs optimal for training and inference. The CPU acts as the general-purpose coordinator. It is responsible for scheduling jobs, managing memory, handling the operating-system work, routing requests, and running the application logic. It coordinates storage, networking, security, virtualization, and I/O. It tells the accelerators (GPU/ASIC) what to do, sends them data, receives their outputs, and manages/monitors the whole process.

正如我们在《互连基础知识 101》中所阐述的，CPU 并不具备处理密集矩阵数学运算的能力，而这正是 GPU 在训练和推理方面表现卓越的原因。CPU 充当通用协调器的角色。它负责调度作业、管理内存、处理操作系统任务、路由请求以及运行应用逻辑。它协调存储、网络、安全、虚拟化和 I/O。它指挥加速器（GPU/ASIC）执行任务，向其发送数据，接收其输出，并管理和监控整个过程。

In a normal chatbot interaction, that surrounding work is secondary, but in an agentic workflow, it becomes central.

在普通的聊天机器人交互中，这类周边工作是次要的；但在智能体工作流中，它则变得至关重要。

For example, in a coding agent, the GPU generates the code, but the CPU runs the shell, opens files, executes tests, reads logs, manages the repository, handles Python or Bash, and decides what should be sent back to the model. For a research agent, the

GPU may summarize and produce the output, but the CPU manages retrieval, search, ranking, document parsing, browser interaction, and API calls. In an enterprise agent, the GPU may reason, but the CPU handles permissions, databases, workflow systems, CRM updates, audit logs, and policy checks.

例如，在编程智能体中，GPU 负责生成代码，而 CPU 则负责运行 Shell、打开文件、执行测试、读取日志、管理代码库、处理 Python 或 Bash 脚本，并决定哪些信息应反馈给模型。对于研究智能体，GPU 可能负责总结并生成输出，但 CPU 则负责管理检索、搜索、排序、文档解析、浏览器交互以及 API 调用。在企业级智能体中，GPU 可能负责推理，而 CPU 则处理权限管理、数据库操作、工作流系统、CRM 更新、审计日志以及策略检查。

Reinforcement learning environments bolster the CPU thesis. When AI systems are trained or evaluated through reinforcement learning, the GPU handles the dense matrix multiplication math, but the CPU often runs the environment that produces the next reward signal. This matters because RL is a repeated loop of action, environment step, and feedback. Depending on the task being RLed, the environment for that task may involve physics simulation, rule-based logic, API calls, database queries, sandboxed code execution, unit tests, web navigation, or multi-agent interaction. These environment steps are difficult to parallelize cleanly on GPUs. As a result, large-scale RL can become constrained by how quickly CPUs can generate experience/data for the model to learn from. The more the industry uses RL to make agents reliable, autonomous, and task-oriented, the more demand rises for the high-core-count CPUs, memory bandwidth, fast storage, and low-latency networking needed to keep GPUs fed with useful training and evaluation data.

强化学习环境进一步支撑了 CPU 论点。当 AI 系统通过强化学习进行训练或评估时，GPU 负责处理密集的矩阵乘法运算，而 CPU 则通常运行产生下一个奖励信号的环境。这一点至关重要，因为强化学习是一个由动作、环境步骤和反馈组成的重复循环。根据所执行的强化学习任务，该任务的环境可能涉及物理模拟、基于规则的逻辑、API 调用、数据库查询、沙箱代码执行、单元测试、网页导航或多智能体交互。这些环境步骤很难在 GPU 上进行干净的并行化。因此，大规模强化学习可能会受限于 CPU 生成供模型学习的经验/数据的速度。业界越是利用强化学习来使智能体变得可靠、自主且以任务为导向，对高核心数 CPU、内存带宽、快速存储以及低延迟网络的需求就越高，因为这些都是确保 GPU 获得充足的有用训练和评估数据所必需的。

Nvidia is building Vera as they want to own the AI factory. **Arm Holdings** (**ARM US**) is building AGI CPU as they want to expand from IP into datacenter silicon. Intel is defending the host CPU layer in Nvidia systems, as they want to preserve the x86 host layer. **AMD** (**AMD US**) is pairing EPYC and Instinct while trying to sell balanced infrastructure. **Qualcomm**'s (**QCOM US**) Snapdragon suite is jockeying to win the on-device orchestration layer within agentic workflows, with the X2 Elite Extreme outperforming benchmarks for multi-threaded productivity tasks. Amazon, Google, and Microsoft want custom silicon control and less dependence on the silicon designers. Each company is taking a different approach, but the direction is the same: the workload is becoming more heterogeneous, and the CPU is becoming more strategic.

Nvidia 正在打造 Vera，因为他们希望掌控 AI 工厂。Arm Holdings (ARM US) 正在开发 AGI CPU，因为他们希望从 IP 领域扩展到数据中心芯片。Intel 正在捍卫 Nvidia 系统中的主机 CPU 层，因为他们希望保留 x86 主机层。AMD (AMD US) 正在将 EPYC 与 Instinct 配对，试图销售平衡的基础设施。Qualcomm (QCOM US) 的 Snapdragon 套件正在竞相赢得代理工作流中的设备端编排层，其 X2 Elite Extreme 在多线程生产力任务的基准测试中表现优异。Amazon、Google 和 Microsoft 则希望掌握定制芯片的控制权，并减少对芯片设计厂商的依赖。每家公司都采取了不同的策略，但方向是一致的：工作负载正变得更加异构化，而 CPU 的战略地位也日益凸显。

The scarce resource moves from just FLOPS to useful work per rack, per watt, per dollar, and per second.

稀缺资源正从单纯的 FLOPS（每秒浮点运算次数）转向每机架、每瓦特、每美元以及每秒所能完成的有效工作。

That means the control/orchestration plane becomes investable. The control plane includes high-core-count CPUs, coherent CPU-GPU interconnects, PCIe and CXL, memory-rich servers, fast storage, low-latency networking, DPUs, secure sandboxes, inference runtimes, orchestration software, vector databases, and observability.

这意味着控制/编排层变得具有投资价值。控制层包括高核心数 CPU、一致性 CPU-GPU 互连、PCIe 和 CXL、高内存服务器、快速存储、低延迟网络、DPU、安全沙箱、推理运行时、编排软件、向量数据库以及可观测性。

Orchestration Layer

PCIe / CXL / Interconnect Chips & IP

Connectivity silicon and design IP that enable host-to-device orchestration, PCIe switching, CXL fabrics, and data movement.

Astera Labs (ALAB US)
PCIe and CXL connectivity chips for AI servers

Credo Technology (CRDO US)
High-speed SerDes, retimers, and connectivity silicon

Microchip Technology (MCHP US)
PCIe switching, timing, and infrastructure connectivity

Rambus (RMBS US)
Interface and memory subsystem IP for data movement

Synopsys (SNPS US)
PCIe, CXL, and interconnect IP plus EDA tools

Cadence Design Systems (CDNS US)
Interconnect IP and design software for system architecture

Arteris (AIP US)
Network-on-chip and system interconnect IP

Montage Technology (6809 HK)
Server interconnect and memory interface silicon

Renesas Electronics (6723 JP)
PCIe timing, switching, and mixed-signal infrastructure chips

Texas Instruments (TXN US)
Signal chain, timing, and connectivity components

Parade Technology (4966 TT)
Interface ICs, redrivers, and retimers

ASPEED Technology (5274 TT)
Server management and infrastructure control silicon

SMART NICS

Programmable network adapters that offload packet processing, virtualization, security, and infrastructure functions from CPUs.

Cisco Systems (CSCO US)
Data center networking and infrastructure offload platforms

Napatech (NAPA NO)
SmartNIC and packet processing acceleration solutions

Silicom (SILC US)
Intelligent NICs and programmable networking adapters

LOW-LATENCY NETWORKS / SWITCHES

High-speed switching, interconnect, and link-enabling components for low-latency AI and cloud networks.

Arista Networks (ANET US)
High-performance Ethernet switching for AI clusters

Cisco Systems (CSCO US)
Enterprise and hyperscale switching platforms

Nokia (NOK US)
Carrier and data center networking infrastructure

Extreme Networks (EXTR US)
High-speed switching and network management

Accton Technology (2345 TT)
White-box and OEM switch hardware platforms

Amphenol (APH US)
High-speed interconnects, cables, and connectors

TE Connectivity (TEL US)
Connectivity components for high-speed networking

Credo Technology (CRDO US)
Connectivity and signal integrity silicon for network links

Citrini Research

In the first AI trade, investors mostly cared about the accelerator supplier. In the agentic trade, they should care about how much actual work a system can complete end-to-end.

在第一波人工智能交易中，投资者主要关注加速器供应商。而在智能体交易中，他们应该关注系统端到端实际能完成多少工作。

There is also a strategic observation here: agents make AI less like a calculator and more like an operating system. Once AI systems are continuously acting, the model becomes one component in a larger control loop. The control loop needs general-purpose compute. It needs to branch, wait, retry, authenticate, serialize, sandbox, fetch, compress, search, and coordinate.

这里还有一个战略性的观察：智能体（agents）使 AI 变得不再像计算器，而更像是一个操作系统。一旦 AI 系统开始持续运行，模型就变成了更大控制回路中的一个组件。这个控制回路需要通用计算。它需要进行分支、等待、重试、身份验证、序列化、沙箱化、抓取、压缩、搜索和协调。

The bear case is that some of today's CPU bottlenecks are software artifacts. The latency stems from Python overhead, poor batching, weak serialization, and immature schedulers, which can be improved. That is true, but the fix for expensive GPU underutilization is better CPU-side scheduling, faster tokenization, cleaner data movement, and more efficient orchestration, then the leverage still exists at the CPU/control plane level. The bottleneck can migrate from hardware to software, but it still remains outside the raw GPU matrix math.

看空者的理由是，当今的一些 CPU 瓶颈是软件层面的产物。延迟源于 Python 的开销、糟糕的分批处理、薄弱的序列化以及不成熟的调度器，而这些都是可以改进的。这固然没错，但如果解决昂贵的 GPU 利用率不足的方法是更好的 CPU 端调度、更快的分词（tokenization）、更简洁的数据移动以及更高效的编排，那么杠杆作用依然存在于 CPU/控制平面层级。瓶颈可能会从硬件迁移到软件，但它仍然存在于原始的 GPU 矩阵运算之外。

Our focus shouldn't simply be on Intel, AMD, and Arm. All three will benefit (and are benefitting) from the agentic AI cycle, but this has already been reflected in their stock

price.

我们的关注点不应仅仅局限于 Intel、AMD 和 Arm。这三家公司都将从代理式 AI（agentic AI）周期中受益（且正在受益），但这已经反映在它们的股价中了。

Before all of you look at the image above and go "yeah, we know" – what the market has just now begun to reprice is the orchestration plane, and this is the primary opportunity.

在你们看到上面的图片并心想"是的，我们早就知道了"之前——请注意，市场才刚刚开始重新定价编排层（orchestration plane），而这正是最主要的机遇所在。



Source: Citrini Research

Positions in basket

as of 2026-05-11 at market close

Ticker	Exchange	Security Type	Industry	Weight	1D Return
APH AMPHENOL CORP-CL A	US	Equity Common Stock	Industrial Electronics | Electronic Connectors	4.00%	-4.34%
EXTR EXTREME NETWORKS INC	US	Equity Common Stock	Communications Telecommunications | Networking Products	4.00%	+1.12%
SILC SILICOM LTD	US	Equity Common Stock	Communications Telecommunications | Networking Products	4.00%	-3.77%
MCHP MICROCHIP TECHNOLOGY INC	US	Equity Common Stock	Technology Semiconductors | Electronic Compo-Semicon	4.00%	-0.06%
NOK NOKIA CORP-SPON ADR	US	Equity ADR	Communications Telecommunications | Networking Products	4.00%	+8.58%
RMBS RAMBUS INC	US	Equity Common Stock	Technology Semiconductors | Electronic Compo-Semicon	4.00%	+4.07%
SNPS SYNOPSYS INC	US	Equity Common Stock	Technology Software | Computer Aided Design	4.00%	-0.06%
TXN TEXAS INSTRUMENTS INC	US	Equity Common Stock	Technology Semiconductors | Electronic Compo-Semicon	4.00%	+3.46%
CDNS CADENCE DESIGN SYS INC	US	Equity Common Stock	Technology Software | Computer Aided Design	4.00%	+0.41%
CSCO CISCO SYSTEMS INC	US	Equity Common Stock	Communications Telecommunications | Networking Products	4.00%	+2.23%
QCOM QUALCOMM INC	US	Equity Common Stock	Technology Semiconductors | Semicon Compo-Intg Circu	4.00%	+8.42%
2345 ACCTON TECHNOLOGY CORP	TT	Equity Common Stock	Communications Telecommunications | Networking Products	4.00%	+9.08%
6723 RENESAS ELECTRONICS CORP	JP	Equity Common Stock	Technology Semiconductors | Semicon Compo-Intg Circu	4.00%	-1.14%
ANET ARISTA NETWORKS INC	US	Equity Common Stock	Communications Telecommunications | Networking Products	4.00%	-3.77%
3533 LOTES CO LTD	TT	Equity Common Stock	Industrial Electronics | Electronic Connectors	4.00%	+5.40%
AIP ARTERIS INC	US	Equity Common Stock	Technology Semiconductors | Electronic Compo-Semicon	4.00%	+3.99%
4966 PARADE TECHNOLOGIES LTD	TT	Equity Common Stock	Technology Semiconductors | Semicon Compo-Intg Circu	4.00%	+9.89%
5274 ASPEED TECHNOLOGY INC	TT	Equity Common Stock	Technology Semiconductors | Electronic Compo-Semicon	4.00%	+5.16%
NAPA NAPATECH A/S	NO	Equity Common Stock	Communications Telecommunications | Networking Products	4.00%	+0.54%
300394 SUZHOU TFC OPTICAL COMMUNI-A	CH	Equity Common Stock	Communications Telecommunications | Telecom Eq Fiber Optics	4.00%	+1.80%
6088 FIT HON TENG LTD	HK	Equity Common Stock	Industrial Electronics | Electronic Compo-Misc	4.00%	+8.59%
CRDO CREDO TECHNOLOGY GROUP HOLDI	US	Equity Common Stock	Communications Telecommunications | Telecommunication Equip	4.00%	+11.52%
ALAB ASTERA LABS INC	US	Equity Common Stock	Technology Semiconductors | Semiconductor Equipment	4.00%	+3.78%
TEL TE CONNECTIVITY PLC	US	Equity Common Stock	Industrial Electronics | Electronic Compo-Misc	4.00%	-0.04%
6809 MONTAGE TECHNOLOGY CO LTD-H	HK	Equity Common Stock	Technology Semiconductors | Electronic Compo-Semicon	4.00%	+11.45%

You can view the basket on Citrindex **here**.

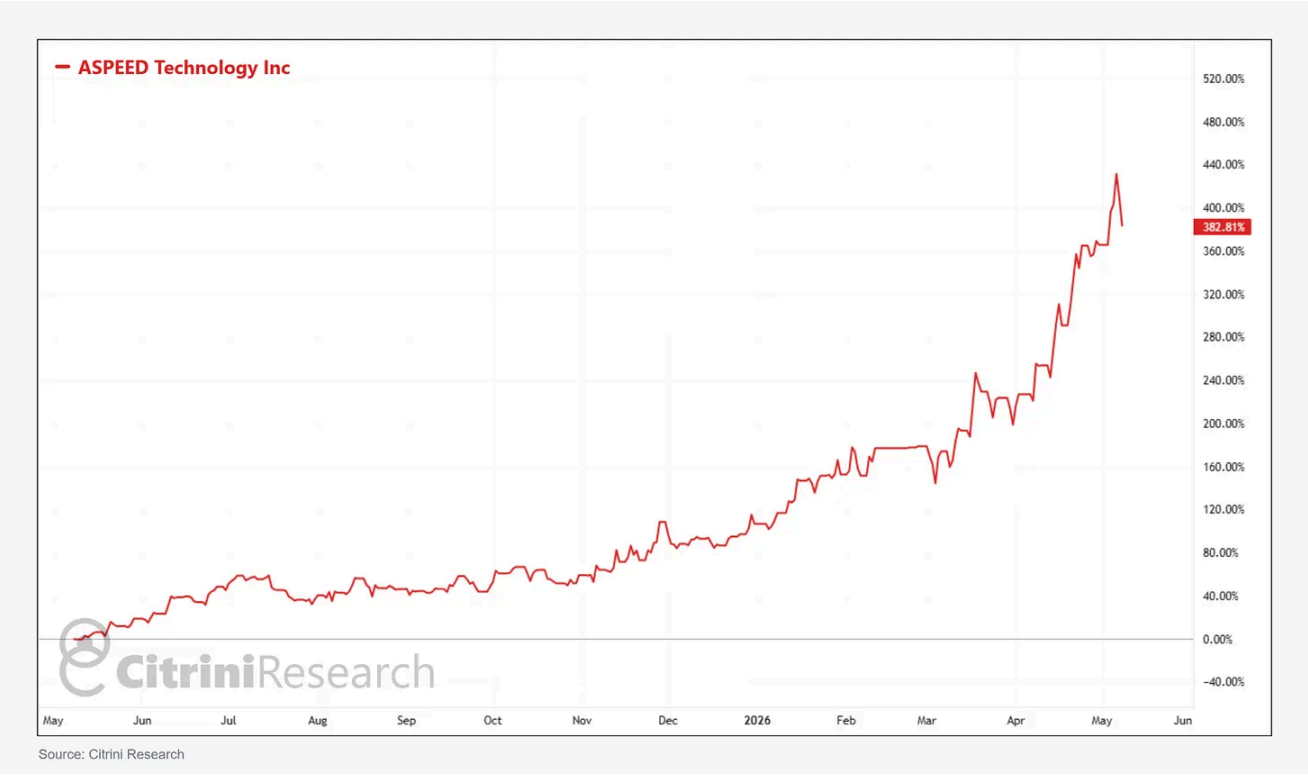
您可以在 Citrindex 上查看该一篮子指数，链接如下。

Here are some specific single stocks to highlight...

以下是一些值得关注的具体个股.....

ASPEED (5274 TT)

信骅科技 ASPEED (5274 TT)



Source: Citrini Research

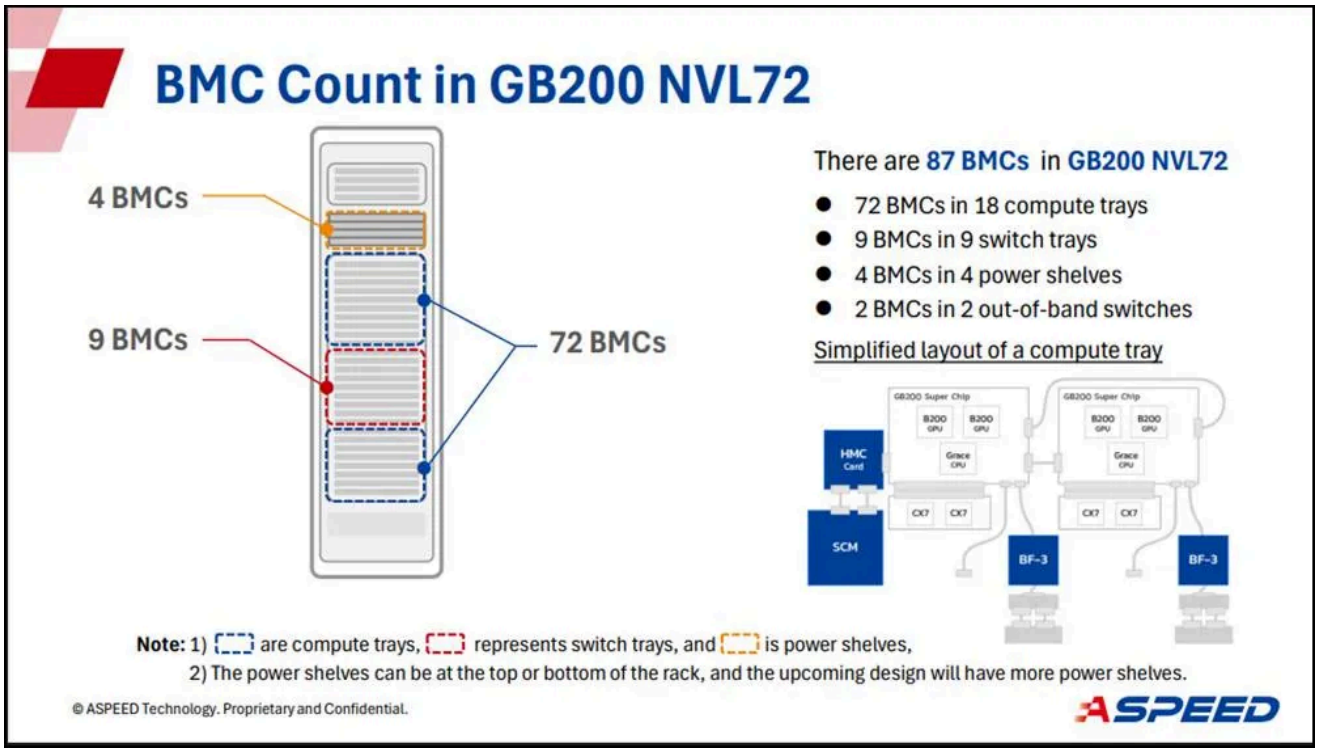
ASPEED is a fabless chip designer known for their BMC (baseboard management controller) chips. A BMC chip is present on a server or rack module that lets operators monitor, control, secure, update, and troubleshoot the system independently of the host CPU/GPU operating system.

ASPEED 是一家以其 BMC（基板管理控制器）芯片而闻名的无晶圆厂芯片设计公司。BMC 芯片存在于服务器或机架模块上，允许运营商独立于主机 CPU/GPU 操作系统对系统进行监控、控制、保护、更新和故障排除。

ASPEED has 80% market share for BMC chips in general CPU servers. They also have 100% market share in BMCs for Nvidia's Hopper/Blackwell servers. Each Blackwell NVL72 has 87 BMC chips, and BMC chip content is expected to grow with the next generation Rubin servers. Liquid cooling is a great driver for ASPEED. Liquid cooling systems embed many liquid detection sensors, requiring BMCs to instantly issue precise power-off commands upon detecting minute amounts of moisture within

milliseconds.

ASPEED 在通用 CPU 服务器的 BMC 芯片市场占有 80% 的份额。在 Nvidia 的 Hopper/Blackwell 服务器 BMC 市场中，他们也拥有 100% 的市场份额。每台 Blackwell NVL72 配备了 87 颗 BMC 芯片，且预计 BMC 芯片的含量将随着下一代 Rubin 服务器的推出而继续增长。液冷技术是 ASPEED 的重要驱动力。液冷系统嵌入了大量液体检测传感器，要求 BMC 在检测到微量水分后的毫秒内，立即发出精确的断电指令。



ASPEED has an ASML-like moat in BMCs. Just like every leading-edge chip is fabbed using ASML's EUV tool, ASPEED's BMC chips are present in x86 CPU servers, Arm-based CPU servers, RISC-V CPU servers, Nvidia GPU racks, AMD GPU racks, Trainium or TPU racks. ASPEED initially forecast BMC TAM at around 46.5 million chips by 2030, but due to growing demand for AI servers and agentic AI, they revised the TAM to 65 million chips by 2030, with BMCs used in AI servers growing 40% in

2027, 35% in 2028, and 30% in 2029–2030.

ASPEED 在 BMC（基板管理控制器）领域拥有类似 ASML 的护城河。正如每一颗尖端芯片都使用 ASML 的 EUV 设备制造一样，ASPEED 的 BMC 芯片也广泛存在于 x86 CPU 服务器、Arm 架构 CPU 服务器、RISC-V CPU 服务器、Nvidia GPU 机架、AMD GPU 机架以及 Trainium 或 TPU 机架中。ASPEED 最初预测到 2030 年 BMC 的市场总量（TAM）约为 4650 万颗，但由于 AI 服务器和代理式 AI（Agentic AI）需求的增长，他们将 2030 年的 TAM 修正为 6500 万颗，其中用于 AI 服务器的 BMC 预计在 2027 年增长 40%，2028 年增长 35%，2029 至 2030 年增长 30%。

We are seeing volume growth at the same time that the ASP of BMC chips is rising as more security and control functions are integrated into them. We can already see the growth from agentic AI demand in their financials. Management had guided Q1 2026 revenue to NT$2.6–2.7 billion with a gross margin of 66.5–67.5% in November 2025. However, the actual Q1 2026 revenue is about NT$3.14 billion, based on the monthly revenue data, which has exceeded that original revenue outlook by roughly 17–21%.

随着越来越多的安全和控制功能被集成到 BMC 芯片中，我们在看到出货量增长的同时，也看到了其平均售价（ASP）的上升。我们已经可以从其财务数据中看到由代理型 AI（Agentic AI）需求带来的增长。管理层曾在 2025 年 11 月指引 2026 年第一季度的营收为 26 亿至 27 亿新台币，毛利率为 66.5% 至 67.5%。然而，根据月度营收数据，2026 年第一季度的实际营收约为 31.4 亿新台币，比最初的营收预期高出约 17% 至 21%。

LOTES (3533 TT)　嘉泽端子 (3533 TT)



Source: Citrini Research

LOTES is a precision connector, CPU sockets, memory connectors, high-speed interconnects, cable assemblies and liquid-cooling hardware manufacturer. They emphasize their vertically integrated production model with in-house design, tooling, molding, stamping, electroplating, and automated assembly. The biggest upside for LOTES comes from their CPU socket sales for Intel, AMD, and Arm platforms, as well as memory connectors, including DDR, SO-DDR and CAMM and Quick Disconnects (QD) for liquid cooling. LOTES has around 35% global market share in CPU sockets.

LOTES 是一家精密连接器、CPU 插槽、内存连接器、高速互连器件、电缆组件及液冷硬件制造商。他们强调其垂直整合的生产模式，涵盖内部设计、模具制造、成型、冲压、电镀和自动化组装。LOTES 最大的增长空间源于其为 Intel、AMD 和 Arm 平台提供的 CPU 插槽销售，以及包括 DDR、SO-DDR、CAMM 在内的内存连接器，以及用于液冷的快速断开接头 (QD)。LOTES 在全球 CPU 插槽市场拥有约 35% 的市场份额。

A key underappreciated driver is not only more CPUs, but more valuable CPU sockets. As next-generation server CPUs increase in pin count, socket complexity, and ASP should rise. Recent company commentary indicated that 2026 CPU sockets average roughly 5,000 pins, with next-generation platforms expected to exceed 7,000 pins. This

creates a content-per-server tailwind even before factoring in unit growth.

一个被低估的关键驱动因素不仅是 CPU 数量的增加，还有更具价值的 CPU 插槽。随着下一代服务器 CPU 引脚数量的增加，插槽的复杂性和平均售价（ASP）也应随之上升。近期公司评论指出，2026 年的 CPU 插槽平均约为 5,000 个引脚，而下一代平台预计将超过 7,000 个引脚。甚至在考虑出货量增长之前，这已经为单台服务器价值量的提升创造了顺风。

At its March 2026 investor meeting, LOTES indicated that 2026 revenue could grow about 15%, with the server business up about 30%, continuing last year's strong server momentum. LOTES reported that Q1 2026 revenue reached NT$9.38 billion, up 21% year over year. This compares favorably with management's full-year revenue growth target of roughly 15%, suggesting that guidance could prove conservative if server momentum continues into the second half.

在 2026 年 3 月的投资者会议上，嘉泽（LOTES）表示 2026 年营收有望增长约 15%，其中服务器业务增长约 30%，延续了去年强劲的服务器增长势头。嘉泽报告称，2026 年第一季度营收达到 93.8 亿新台币，同比增长 21%。这一表现优于管理层设定的全年营收增长约 15% 的目标，表明如果服务器业务的增长势头持续到下半年，该业绩指引可能会显得保守。

We believe that they will beat that target and revenue growth will accelerate, especially by Q3/Q4. The product mix is also shifting more towards servers, away from consumer products, which will be favorable for their growth. Because of raw material costs, especially due to higher copper, gold, and resin prices, the company has started negotiating price increases with its customers. This is a big break from the past, as LOTES used to make small price cuts every quarter for their customers. Management also stated that server order visibility had improved from 4-6 weeks last year to 12 weeks to 4 months. LOTES is also expanding outside China, which may become increasingly important for hyperscale and server customers seeking supply-chain diversification. Factories in Vietnam were responsible for 16% of revenue in 2025, with the management targeting 20%-25% this year. More than 70% of Vietnam's capacity is focused on server products. This gives LOTES a strategic advantage if customers

require non-China capacity for AI-server supply chains.

我们相信他们将超越这一目标，且营收增长将进一步加速，特别是在第三和第四季度。产品组合也正更多地从消费级产品转向服务器，这将有利于其增长。由于原材料成本（特别是铜、金和树脂价格上涨）的影响，公司已开始与客户协商提价。这与过去相比是一个重大突破，因为 LOTES 此前通常每季度都会为客户提供小幅降价。管理层还表示，服务器订单的可见性已从去年的 4-6 周提高到 12 周至 4 个月。此外，LOTES 正在向中国以外地区扩张，这对于寻求供应链多元化的超大规模数据中心和服务器客户来说可能变得日益重要。越南工厂在 2025 年贡献了 16% 的营收，管理层今年的目标是达到 20%-25%。越南 70% 以上的产能集中在服务器产品上。如果客户在 AI 服务器供应链中需要非中国产能，这将为 LOTES 带来战略优势。

FIT Hon Teng (6088 HK)

鸿腾精密 (6088 HK)



Source: Citrini Research

Foxconn Interconnect Technology, or **FIT Hon Teng** (**6088 HK**), is the interconnect and precision-component arm of the Hon Hai/Foxconn group. Like LOTES, FIT supplies CPU sockets, memory sockets, DDR5 DIMM connectors, PCIe CEM connectors, M-CRPS power connectors, high-speed internal cables, and backplane products. FIT has 40% market share in CPU sockets. The upside for FIT comes from

their CPU socket sales, high-speed 224G+/448G connectivity, optical interconnects, cabinet connectors, and high-speed cables.

鸿腾精密科技（FIT Hon Teng，股票代码：6088 HK）是鸿海/富士康集团旗下的互连解决方案和精密组件部门。与嘉泽（LOTES）类似，鸿腾精密供应 CPU 插槽、内存插槽、DDR5 DIMM 连接器、PCIe CEM 连接器、M-CRPS 电源连接器、高速内部线缆以及背板产品。鸿腾精密在 CPU 插槽市场拥有 40% 的市场份额。该公司的增长潜力源于其 CPU 插槽销售、高速 224G+/448G 连接技术、光互连、机柜连接器以及高速线缆业务。

FIT's long-term guidance shows Cloud/Data Center revenue mix rising from 16% in 2025 to the mid-20% range in 2026, high-20% range in 2027, and low-30% range by 2028. ***We expect them to easily beat this forecast.***

鸿腾精密（FIT）的长期指引显示，云端/数据中心业务的营收占比将从 2025 年的 16% 上升至 2026 年的 25% 左右，2027 年达到 25%-30% 之间，并到 2028 年达到 30% 出头的水平。我们预计他们将轻松超越这一预测。

Montage Technology (6809 HK)

澜起科技 (688001 SH)



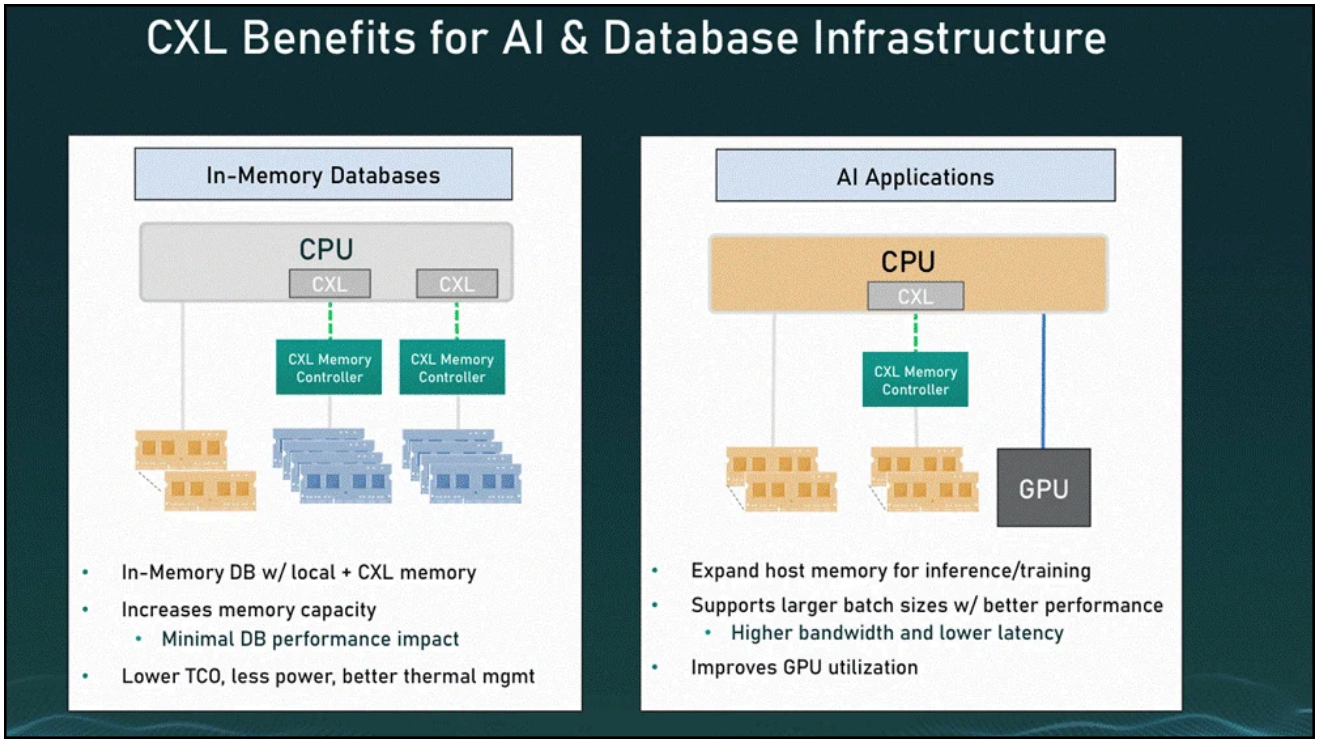
Source: Citrini Research

Montage Technology is a fabless semiconductor design company focused on high-speed interconnect chips for cloud computing and AI infrastructure. They primarily design and sell memory-interface chips, PCIe/CXL interconnect chips, CXL memory expansion controllers, and clock chips. The company is positioning itself as a “content per server” beneficiary since higher demand for agentic AI workloads requires more CPU servers, more memory capacity and higher memory bandwidth, demand rises for DDR5 server modules and the interface chips used on those modules.

澜起科技（Montage Technology）是一家专注于云计算和 AI 基础设施高速互连芯片的无晶圆厂半导体设计公司。他们主要设计并销售内存接口芯片、PCIe/CXL 互连芯片、CXL 内存扩展控制器以及时钟芯片。该公司将自己定位为“单服务器价值量”增长的受益者，因为代理型 AI 工作负载的更高需求需要更多的 CPU 服务器、更大的内存容量和更高的内存带宽，从而带动了对 DDR5 服务器内存模组及其所使用的接口芯片的需求。

The first growth driver for Montage is higher DDR5 content per server. Every CPU server requires memory modules, and Agentic AI requires more host memory alongside the XPU’s HBM. Montage benefits because they design RCD, DB, MRCD/MDB, PMIC, which are present on server memory modules and help CPUs access memory faster in a more reliable manner. The second driver is rising silicon content per module. As CPU core counts and memory speeds increase, signal-integrity challenges become harder, increasing the importance of interface chips.

澜起科技（Montage）的第一大增长驱动力是单台服务器中更高的 DDR5 含量。每台 CPU 服务器都需要内存模组，而代理式人工智能（Agentic AI）在 XPU 的 HBM 之外，还需要更多的宿主内存。澜起科技从中受益，因为他们设计 RCD、DB、MRCD/MDB 以及 PMIC，这些芯片存在于服务器内存模组上，旨在帮助 CPU 以更可靠的方式更快地访问内存。第二个驱动力是单个模组中硅含量的提升。随着 CPU 核心数量和内存速度的增加，信号完整性挑战变得愈发严峻，从而提升了接口芯片的重要性。



This is especially relevant for newer high-bandwidth memory-module architectures such as MRDIMM/MCRDIMM, where Montage sells MRCD and MDB chipsets in addition to conventional DDR5 interface products. The third growth driver is PCIe/CXL connectivity. AI servers require high speed links between CPUs, GPUs, accelerators, SSDs and NICs. Montage's PCIe Retimers compensate for channel attenuation, reduce jitter, and extend high-speed signal reach in servers, storage systems, and accelerators. Its CXL MXC chips provide another growth vector by enabling memory expansion, pooling and sharing across hosts, which can improve memory utilization in data-intensive AI workloads.

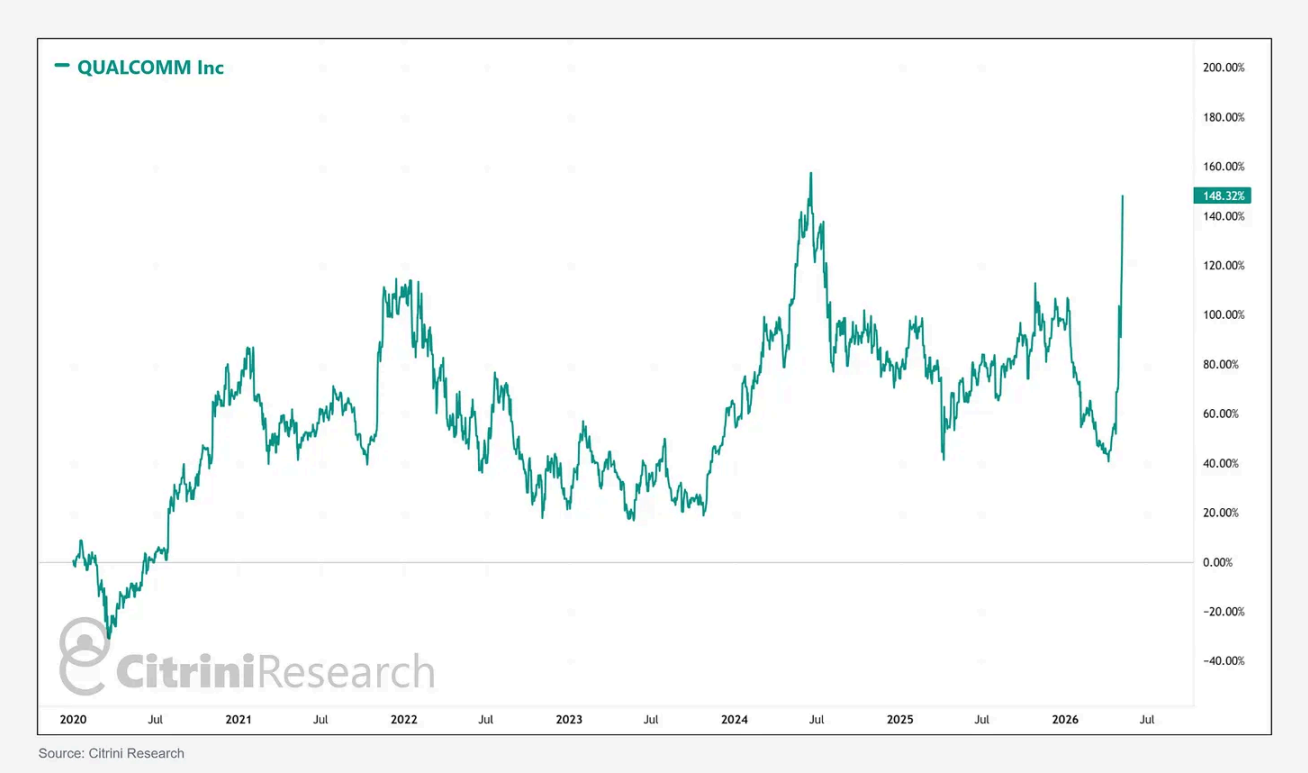
这对于 MRDIMM/MCRDIMM 等新型高带宽内存模组架构尤为重要，在这些架构中，澜起科技（Montage）除了销售传统的 DDR5 接口产品外，还销售 MRCD 和 MDB 芯片组。第三个增长驱动力是 PCIe/CXL 连接。AI 服务器需要 CPU、GPU、加速器、SSD 和网卡之间的高速链路。澜起科技的 PCIe Retimer 芯片可补偿通道衰减、减少抖动，并扩展服务器、存储系统和加速器中的高速信号传输距离。其 CXL MXC 芯片通过实现跨主机的内存扩展、池化和共享，提供了另一个增长维度，从而提高数据密集型 AI 工作负载中的内存利用率。

Montage was the largest supplier of memory interconnect chips in 2024 and held a 36.8% market share. Q1 2026 was exceptionally strong for Montage, gross margin reached 69.8%, up 9.3% YoY, driven by strong shipments of DDR5 RCD chips and new interconnect products. Interconnect-chip gross margin reached 71.5%, up 7.0% YoY.

Gross profit reached RMB1.019 billion, up 38.0%.

澜起科技（Montage）是 2024 年最大的内存接口芯片供应商，占据了 36.8% 的市场份额。2026 年第一季度对澜起科技而言表现异常强劲，在 DDR5 RCD 芯片和新型互连产品强劲出货的推动下，毛利率达到 69.8%，同比增长 9.3%。互连芯片毛利率达到 71.5%，同比增长 7.0%。毛利达到 10.19 亿元人民币，增长 38.0%。

Qualcomm (QCOM US) 高通 (QCOM US)



Source: Citrini Research

Our pet favorite for 2026 has been Qualcomm, and it has certainly tested our patience. The stock was a let-down for the beginning of the year, trapped in a narrative treating it as a mature handset company with Chinese headwinds and fading licensing revenues. Now, however, the market is beginning to see a company that represents a truly asymmetric bet. One of the cheapest names in the semiconductor index is now positioned to re-rate aggressively as a potential CPU winner with an on-device kicker.

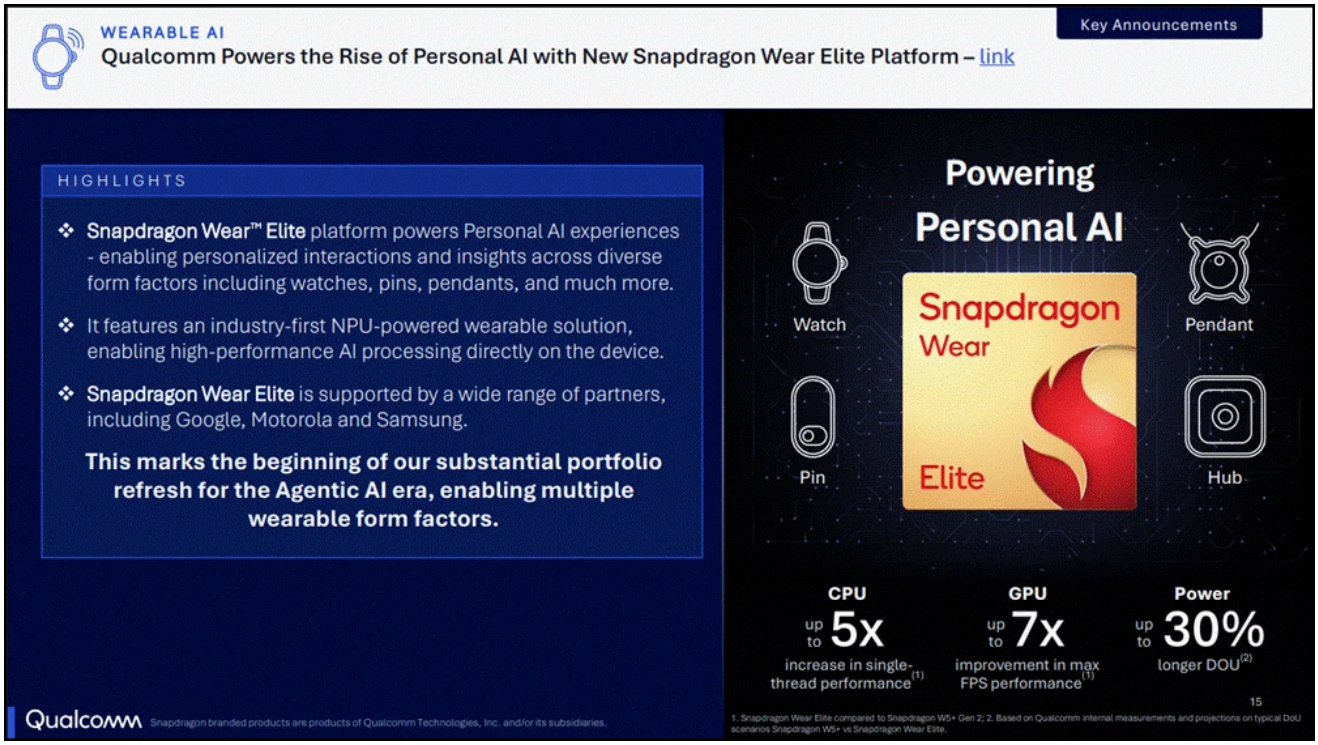
我们对 2026 年最看好的标的一直是 Qualcomm，而它确实考验了我们的耐心。今年年初，该股表现令人失望，受困于将其视为一家面临中国市场逆风且授权收入逐渐衰减的成熟手机公司的叙事中。然而，现在市场开始看到一家代表着真正不对称博弈的公司。作为半导体指数中最便宜的名字之一，它现在有望凭借其作为潜在 CPU 赢家的地位以及端侧 AI 的加持，迎来激进的估值重塑。

The bull case rests on two re-rating catalysts that the consensus is not pricing. The first is the data center CPU opportunity. Qualcomm has been building toward this for years through the **Nuvia** acquisition and the **Oryon** CPU core, and the company is now positioned as a credible alternative to the Arm-Intel-AMD triopoly for hyperscaler custom silicon. There is essentially zero data center revenue in the current model and zero data center revenue in the sell-side estimates.

看涨逻辑基于两个尚未被市场共识定价的估值重估催化剂。首先是数据中心 CPU 的机遇。高通多年来通过收购 Nuvia 和研发 Oryon CPU 核心一直在为此做准备，公司目前已定位为超大规模计算中心定制芯片领域中，除 Arm-Intel-AMD 三巨头之外的一个可靠替代方案。在当前的财务模型和卖方预测中，数据中心业务的营收贡献基本为零。

The second and more contentious catalyst is on-device inference. As HBM supply tightness pushes memory prices higher, the economics of running low-complexity, repetitive inference workloads in hyperscale data centers progressively deteriorate. This does not mean that all inference will leave the cloud. Large reasoning, long-context, and agentic workflows will remain in the data center. But workloads where low latency, privacy, and repetitive user interaction matter face a growing economic incentive to migrate to the device.

第二个且更具争议的催化剂是端侧推理。随着 HBM 供应紧张推高内存价格，在超大规模数据中心运行低复杂度、重复性推理工作负载的经济效益正逐步恶化。这并不意味着所有推理都会离开云端。大型推理、长上下文和智能体（agentic）工作流仍将留在数据中心。但对于那些注重低延迟、隐私和重复性用户交互的工作负载，将其迁移到设备端的经济激励正日益增强。



This is also a pattern that has played out repeatedly across compute cycles. Early on, infrastructure and capital concentrate on the server side. Over time, as costs come down, software matures, and user experience becomes the binding constraint, that compute gradually diffuses toward the consumer edge. From mainframe to personal PC, from server-centric computing to mobile, and now from cloud AI to on-device AI.

这也是在计算周期中反复出现的模式。早期，基础设施和资本集中在服务器端。随着时间的推移，成本下降，软件趋于成熟，用户体验成为核心约束，计算能力逐渐向消费者终端扩散。从大型机到个人电脑，从以服务器为中心的计算到移动端，现在正从云端 AI 转向端侧 AI。

Qualcomm holds the platform best positioned to capture this transition directly. Its low-power compute footprint — already embedded across smartphones, PCs, automotive, and edge devices — combined with its integrated NPU/CPU/GPU architecture, places it at the center of any redistribution of cloud inference back to the device. We do not know precisely when this shift inflects. But as memory price inflation and AI usage growth compound simultaneously, on-device inference moves from optionality toward inevitability. And at that point, Qualcomm will no longer

trade at today's valuation.

高通拥有最有利的平台地位，能够直接捕捉这一转型。其低功耗计算足迹已嵌入智能手机、个人电脑、汽车和边缘设备中，结合其集成的 NPU/CPU/GPU 架构，使其处于云端推理向设备端回流的中心。我们无法准确预知这一转变何时发生拐点。但随着内存价格通胀和 AI 使用量增长的同步叠加，端侧推理正从一种可选项演变为必然趋势。届时，高通的估值将不再处于今天的水平。

Neoclouds: The Inference Shortage

Neoclouds：推理算力短缺

As AI becomes a work platform, inference becomes a recurring production workload. A developer using Claude Code or Codex is running a loop. The same applies for agentic research, legal review, financial analysis, customer support, sales ops, cybersecurity, and eventually – most knowledge work.

随着 AI 成为一种工作平台，推理正演变为一种经常性的生产负载。开发者在使用 Claude Code 或 Codex 时实际上是在运行一个循环。同样的情况也适用于智能体化研究、法律审查、财务分析、客户支持、销售运营、网络安全，以及最终——绝大多数的知识性工作。

The question today is: "**Who has enough live capacity to serve millions of agents while users are actually trying to get work done?**"

当今的问题是："当用户真正尝试完成工作时，谁拥有足够的实时算力来服务数百万个智能体？"

The shortage is not theoretical. From 2023 to 2025, OpenAI's available compute went from roughly 0.2 GW to 1.9 GW. Revenue moved from $2 billion ARR to more than $20 billion over the same period. More importantly, OpenAI said more compute during that period would have led to faster adoption and monetization, and called compute

the scarcest resource in AI.

这种短缺并非理论上的。从 2023 年到 2025 年，OpenAI 的可用算力从大约 0.2 GW 增长到 1.9 GW。同期，其年经常性收入（ARR）从 20 亿美元增长到超过 200 亿美元。更重要的是，OpenAI 表示，在此期间更多的算力本可以带来更快的采用和变现速度，并称算力是 AI 领域最稀缺的资源。

The usage data underneath this is even more important than the headline revenue numbers. By March 2026, OpenAI was generating $2 billion in revenue per month (up to $3 billion as of May). Enterprise had become more than 40% of revenue. APIs were processing more than 15 billion tokens per minute. Codex had more than 2 million weekly users, up 5x in three months, with usage growing more than 70% month over month.

这些数据背后的使用情况比头条营收数字更为重要。到 2026 年 3 月，OpenAI 每月的营收已达 20 亿美元（截至 5 月已增至 30 亿美元）。企业级业务已占营收的 40% 以上。API 每分钟处理的 token 数量超过 150 亿个。Codex 的周活跃用户超过 200 万，在三个月内增长了 5 倍，且使用量每月增长超过 70%。

Likewise, Anthropic's business subscriptions to Claude Code quadrupled from the start of 2026 and enterprise use now represents more than half of Claude Code's revenue. Anthropic also said customers spending over $1 million annually on Claude doubled from more than 500 to more than 1,000 in less than two months.

同样，Anthropic 的 Claude Code 商业订阅量自 2026 年初以来翻了两番，企业级使用目前占 Claude Code 营收的一半以上。Anthropic 还表示，在不到两个月的时间里，每年在 Claude 上支出超过 100 万美元的客户数量翻了一番，从 500 多家增加到 1,000 多家。

Anthropic's run-rate revenue reached $45 billion by May 2026, up 400% from about $9 billion at the end of 2025. They also admitted that the growth had strained infrastructure and hurt reliability/performance during peak hours.

到 2026 年 5 月，Anthropic 的年化营收达到 450 亿美元，较 2025 年底的约 90 亿美元增长了 400%。他们还承认，这种增长使基础设施承受了巨大压力，并在高峰时段损害了可靠性和性能。

AI labs need usable compute now.

AI 实验室现在需要可用的算力。

Anthropic signed a deal with Amazon for up to 5 GW of compute, including nearly 1 GW by the end of 2026. It signed a deal with Google and Broadcom for multiple gigawatts of TPU capacity coming online starting in 2027 and committed to spending $200 billion on Google Cloud over the next 5 years. Anthropic then announced a SpaceX agreement for more than 300 MW of capacity and over 220,000 Nvidia GPUs within a month. Anthropic also signed a $1.8 billion, 7-year cloud deal with Akamai for their distributed cloud infra to serve agents at low latency.

Anthropic 与亚马逊签署了一项高达 5 GW 算力的协议，其中包括到 2026 年底前投入使用的近 1 GW 算力。它还与谷歌和博通签署了协议，计划从 2027 年开始上线数个吉瓦（GW）规模的 TPU 产能，并承诺在未来 5 年内向谷歌云投入 2000 亿美元。随后，Anthropic 在一个月内宣布了与 SpaceX 的协议，涉及超过 300 MW 的容量和超过 220,000 颗英伟达 GPU。此外，Anthropic 还与 Akamai 签署了一项为期 7 年、价值 18 亿美元的云服务协议，利用其分布式云基础设施为智能体（agents）提供低延迟服务。

Once Anthropic got access to SpaceX capacity, it immediately raised Claude Code limits, removed peak-hour reductions for paid users, and increased Claude Opus API limits. More compute means more product availability, more usage, and more monetization.

一旦 Anthropic 获得了 SpaceX 的算力资源，它立即提高了 Claude Code 的限制，取消了付费用户的尖峰时段限制，并提升了 Claude Opus 的 API 限制。更多的算力意味着更高的产品可用性、更多的使用量以及更强的变现能力。

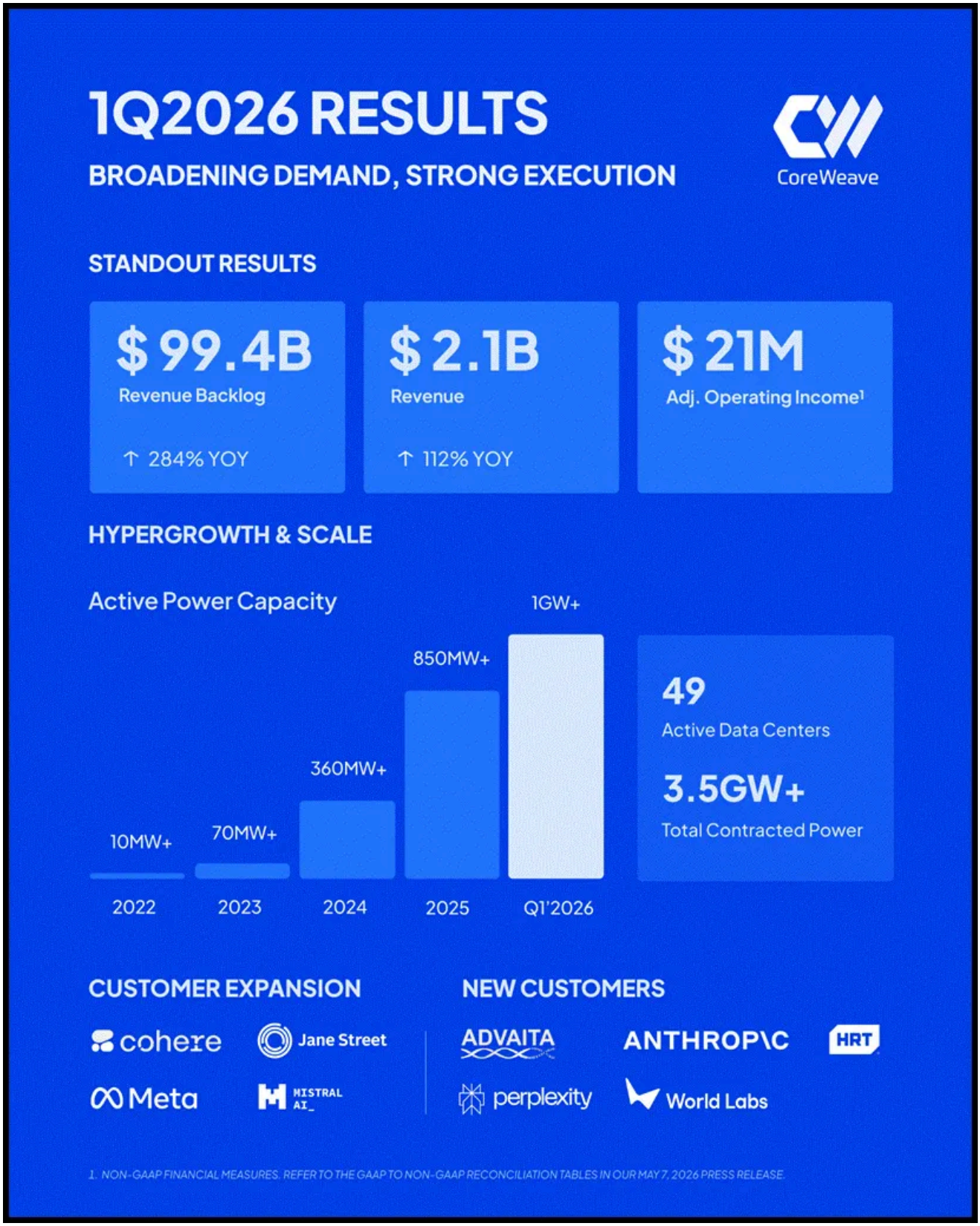
Anthropic is grabbing capacity across every credible route because demand growth is running ahead of infrastructure. If Claude Code, Codex, or a workspace agent is being used inside a real workflow, the customer expects the model to respond immediately. That forces AI labs to overprovision for peaks, maintain redundancy, and keep spare serving capacity online just like an electricity grid. This is why a shortage shows up as throttled users, lower rate limits, elevated errors, and degraded product experience rather than just “higher compute costs.” **If you are a neocloud with live capacity, then**

your product is a relief valve in this environment.

Anthropic 正在通过所有可靠途径争夺产能，因为需求的增长速度已经超过了基础设施的建设速度。如果 Claude Code、Codex 或工作区代理被应用于实际工作流中，客户会期望模型能够立即响应。这迫使 AI 实验室必须像管理电网一样，为高峰时段过度配置资源、保持冗余并维持在线备用服务能力。这就是为什么供应短缺表现为用户受限、速率限制降低、错误率升高以及产品体验下降，而不仅仅是“计算成本增加”。如果你是一家拥有现成产能的新型云服务商（neocloud），那么在这种环境下，你的产品就是泄压阀。

CoreWeave (**CRWV US**) reported $66.8 billion of revenue backlog at year-end 2025. They had $60.7 billion of remaining performance obligations, with committed contracts carrying a weighted-average duration of about five years. By Q1 2026, RPO grew to $99.4 billion, and active power capacity reached 1GW+. They have announced a $21 billion Meta expansion through 2032, a multi-year Anthropic agreement with undisclosed financial terms, and a $6 billion Jane Street AI cloud commitment plus a $1 billion equity investment. **Hut 8** (**HUT US**) just signed a 15-year AI data-center lease in Texas worth $9.8 billion for an initial 352 MW phase. The important details were in the structure: take-or-pay, triple-net, and no early termination.

CoreWeave (CRWV US) 报告称，截至 2025 年底，其营收积压订单达 668 亿美元。公司剩余履约义务（RPO）为 607 亿美元，已签署合同的加权平均期限约为五年。到 2026 年第一季度，RPO 增长至 994 亿美元，运行功率容量达到 1GW 以上。他们已宣布了一项到 2032 年总额 210 亿美元的 Meta 扩建计划、一项财务条款未披露的 Anthropic 多年期协议，以及来自 Jane Street 的 60 亿美元 AI 云承诺和 10 亿美元股权投资。Hut 8 (HUT US) 刚刚在德克萨斯州签署了一份为期 15 年、价值 98 亿美元的 AI 数据中心租赁协议，用于初始 352 MW 阶段。其关键细节在于合同结构：采取“照付不议”（take-or-pay）、绝对净租赁（triple-net）且不可提前终止。



The Trade 交易

We should not underwrite permanent GPU scarcity. But the explosion in inference demand in the last several quarters, supported by these highly visible deals, suggests that anyone who has available capacity now can re-rate higher – especially names that

we wouldn't necessarily consider buy-and-hold bluechips. The mad scramble for compute could lead to a more speculative catch-up trade in a wide range of lower-tier operators – including bitcoin mining conversions.

我们不应预设 GPU 会永久稀缺。但过去几个季度推理需求的爆发，在这些备受瞩目的交易支持下，表明任何目前拥有可用产能的公司都可以获得更高的估值重估——尤其是那些我们不一定认为是值得长期持有的蓝筹股。对算力的疯狂争夺可能会引发更广泛的低端运营商（包括转型中的比特币挖矿公司）出现更具投机性的补涨交易。

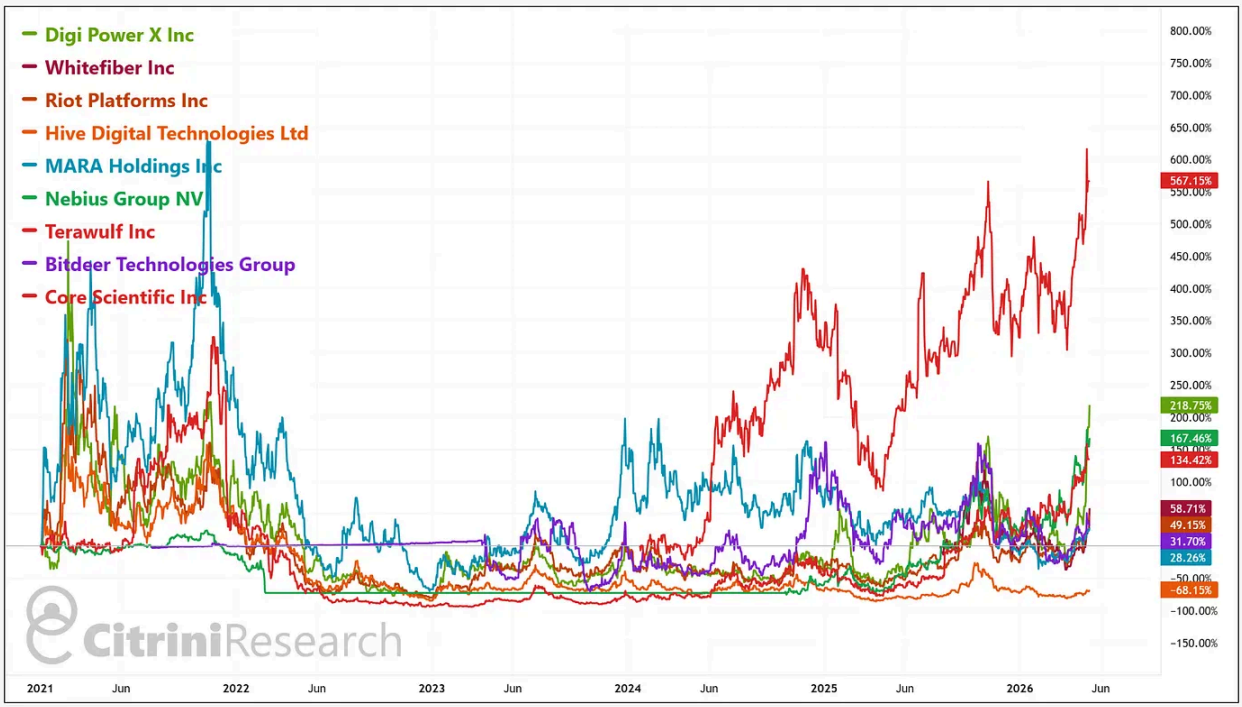
Sellable / Uncommitted AI Capacity

Company	Sellable MW	Timeline	Capacity
1. **Core Scientific** (NASDAQ: CORZ)	~520–685 MW	2026–2027	Broad U.S. site base, with Pecos, Texas standing out as the main AI conversion site. The explicit unleased capacity disclosure makes Core one of the clearest merchant opportunities.
2. **Bitdeer** (NASDAQ: BTDR)	225 MW core + 99 MW possible	Q4 2026–Q1 2027	Tydal, Norway is the main near-term opportunity at 225 MW. There is also possible follow-on capacity from Knoxville, Tennessee and Wenatchee, Washington.
3. **TeraWulf** (NASDAQ: WULF)	~480 MW late-window + 50–138 MW Lake Hawkeye	Beginning 2027 / likely H2 2027	Hawesville, Kentucky and Lake Hawkeye in Lansing, New York look like the main merchant opportunities, especially for 2027 capacity.
4. **Keel Infrastructure** (NASDAQ: KEEL)	~178 MW countable + 300 MW upside	Now–H1 2027	The near-term pipeline includes Moses Lake, Washington; Sharon, Pennsylvania; and Panther Creek, Pennsylvania, with more room to grow into 2027.
5. **HIVE Digital Technologies** (NASDAQ: HIVE)	~8 MW immediate + 7.6 MW in 2027 + 50 MW Grand Falls	Immediate through 2027	Immediate capacity comes from British Columbia and Manitoba, while the larger 50 MW conversion opportunity is at Grand Falls, New Brunswick.
6. **Riot Platforms** (NASDAQ: RIOT)	Potentially hundreds of MW, not yet countable	2026–2027	Rockdale, Texas and Corsicana, Texas offer real upside, but the company has not given a clean schedule for sellable AI-ready capacity.
7. **Nebius Group** (NASDAQ: NBIS)	~120 MW France by YE26 + 310 MW Finland project	Late 2026–2027	Béthune, France could provide roughly 120 MW by year-end 2026, while Lappeenranta, Finland is a 310 MW project that appears open rather than tied to a single customer.

Citrini Research

Two additional neoclouds that likely will outperform in the wake of the Cerebras IPO due to existing contracts with them are **WhiteFiber** (**WYFI US**) and **Digi Power X** (**DGXX US**). They've begun to take off, but it's not hard to see how an extremely successful IPO for CBRS can result in a much stronger view being taken on these.

由于与 Cerebras 签订了现有合同，另外两家新云服务商 WhiteFiber (WYFI US) 和 Digi Power X (DGXX US) 很可能在 Cerebras IPO 后表现优异。它们已经开始发力，但不难预见，如果 CBRS 的 IPO 取得巨大成功，市场对这两家公司的看法将会变得更加积极。



Source: Citrini Research



Source: Citrini Research

Positions in basket — as of 2026-05-11 at market close

Ticker	Exchange	Security Type	Industry	Weight	1D Return
DGXX DIGI POWER X INC	US	Equity Common Stock	Financial Investment Companies | Investment Companies	13.50%	+12.07%
WYFI WHITEFIBER INC	US	Equity Common Stock	Technology Software | Computer Software	13.50%	+26.96%
RIOT RIOT PLATFORMS INC	US	Equity Common Stock	Financial Investment Companies | Investment Companies	10.43%	+5.23%
HIVE HIVE DIGITAL TECHNOLOGIES LT	US	Equity Common Stock	Financial Investment Companies | Investment Companies	10.43%	+6.74%
MARA MARA HOLDINGS INC	US	Equity Common Stock	Financial Investment Companies | Investment Companies	10.43%	+3.48%
NBIS NEBIUS GROUP NV	US	Equity Common Stock	Technology Software | Computer Software	10.43%	+5.11%
WULF TERAWULF INC	US	Equity Common Stock	Financial Investment Companies | Investment Companies	10.43%	-0.09%
BTDR BITDEER TECHNOLOGIES GROUP-A	US	Equity Common Stock	Financial Investment Companies | Investment Companies	10.43%	-2.23%
CORZ CORE SCIENTIFIC INC	US	Equity Common Stock	Financial Investment Companies | Investment Companies	10.43%	+0.13%

You can view the basket on Citrindex here.

您可以在 Citrindex 上查看该一篮子指数，链接如下。

AI Materials Bottlenecks AI 材料瓶颈

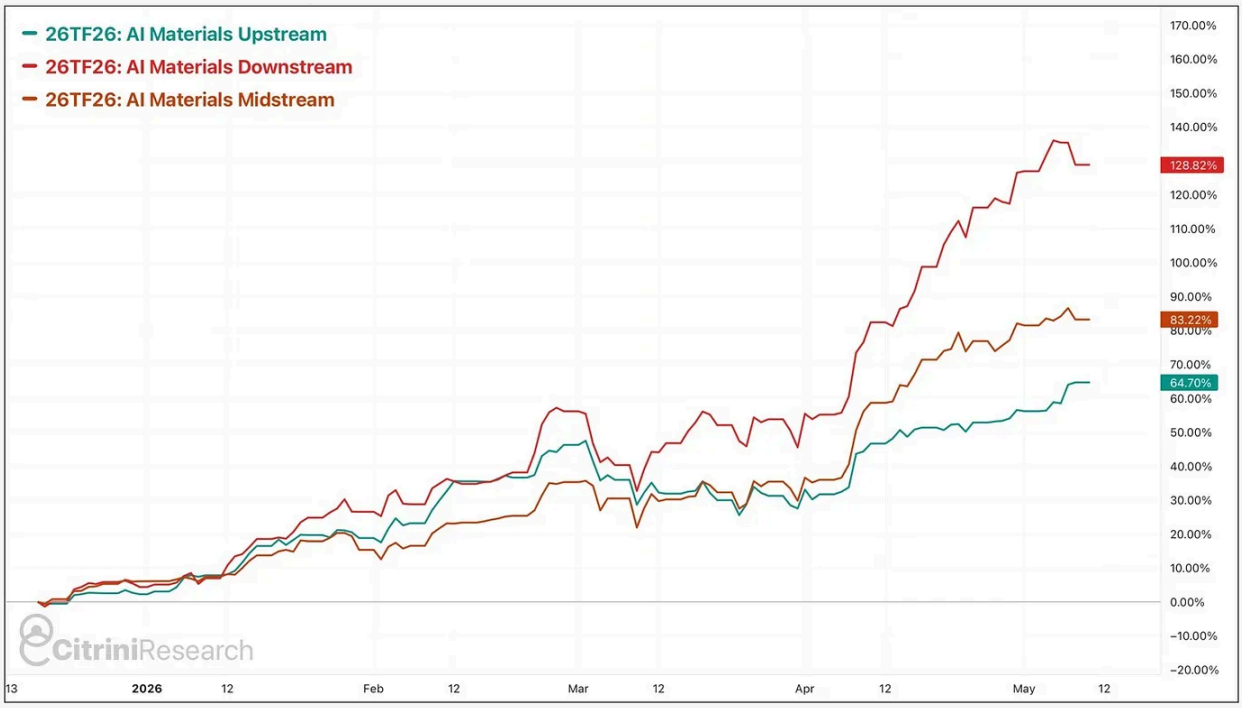
In Trade #5 of **26 Trades for 2026**, we detailed our view that the AI trade would unavoidably broaden out into the previously-viewed-as-commoditized layer that we dubbed “AI Materials”. We split this into 3 segments – **upstream** (resins, coatings, polymers, films, powders, chemicals), **midstream** (laminates, foil, glass cloth, tooling) and **downstream** (printed circuit boards, IC substrates, system-level assembly).

在“2026 年的 26 笔交易”中的第 5 笔交易里，我们详细阐述了我们的观点：AI 交易将不可避免地扩展到此前被视为商品化的层级，我们将其称为“AI 材料”。我们将这一领域分为三个细分板块——上游（树脂、涂层、聚合物、薄膜、粉末、化学品）、中游（覆铜板、箔、玻璃布、模具）和下游（印刷电路板、IC 载板、系统级组装）。

Let’s take a look at how they’ve done.

让我们来看看他们的表现如何。

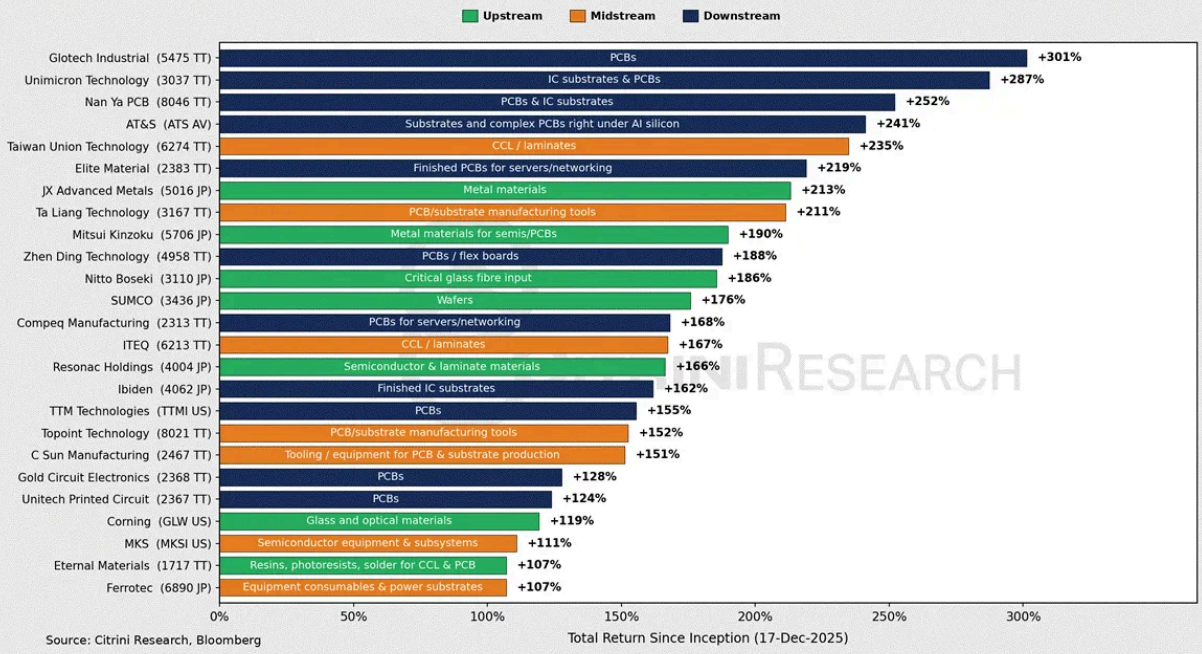
AI Materials Baskets: Performance Since Inception



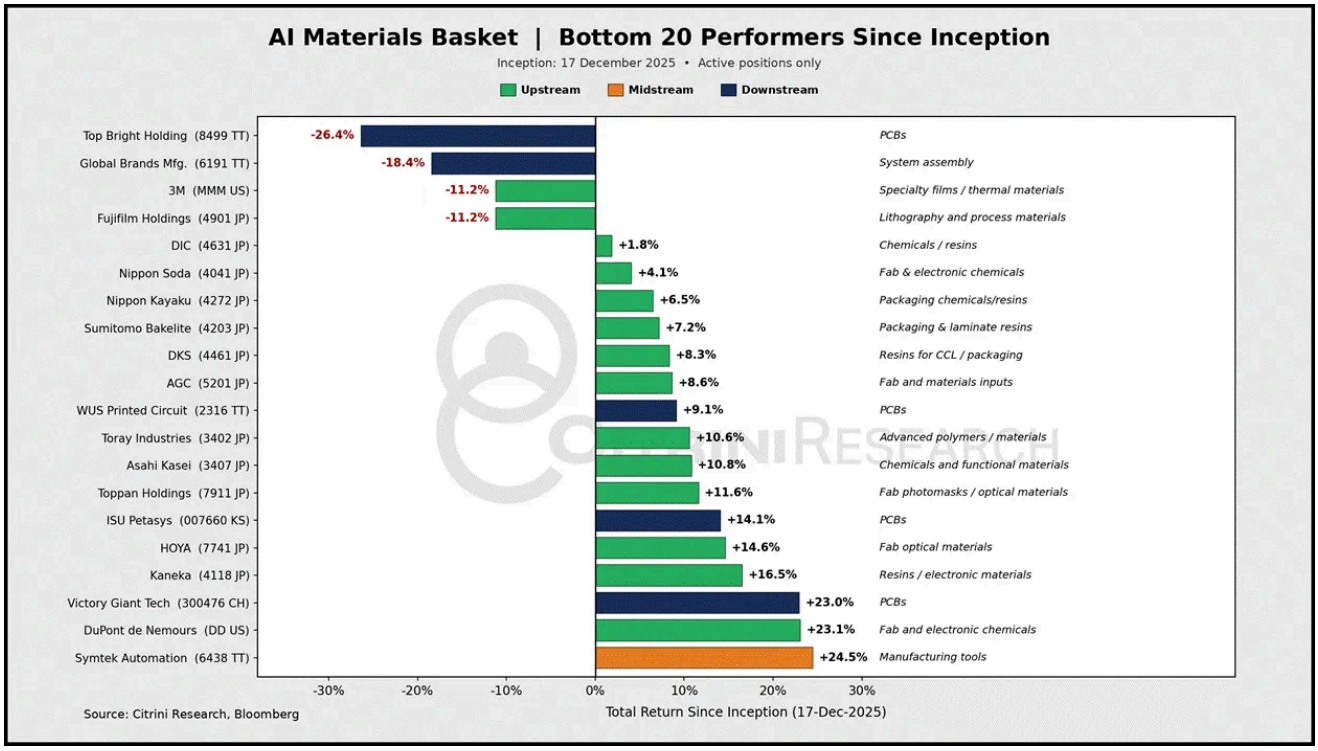
Source: Citrini Research

AI Materials Basket | Top 25 Performers Since Inception

Inception: 17 December 2025 • Active positions only



Source: Citrini Research, Bloomberg



All three of these baskets have performed well, following our expectation and historical precedent that the closer they were to the silicon the better they'd do.

这三个组合的表现都非常出色，符合我们的预期和历史先例，即越接近硅端，表现就越好。

A quick recap of how we viewed the status of AI material bottlenecks back in December: Tier 1 was already-running bottlenecks (T-glass, ABF film, CCL, glass cloth), Tier 2 was inflecting (ceramic substrates, HVLP copper foil, low-loss CCL grades), and Tier 3 was the watchlist of materials where demand was inflecting but supply hadn't visibly tightened. Our top pick for the name with the most exposure to as many potential bottleneck areas was **Resonac** (**4004 JP**), and it still remains well positioned to capture further.

简要回顾一下我们在去年 12 月对 AI 材料瓶颈状态的看法：第一梯队是已经出现的瓶颈（T 玻璃、ABF 薄膜、CCL、玻璃纤维布），第二梯队是正处于拐点的材料（陶瓷基板、HVLP 铜箔、低损耗 CCL 等级），第三梯队是观察名单，即需求正在发生拐点但供应尚未明显收紧的材料。我们认为在尽可能多潜在瓶颈领域中暴露程度最高的首选标的是 Resonac (4004 JP)，且它目前仍处于有利地位，有望进一步获益。

All of our severe bottlenecks have gotten worse. T-Glass capacity relief has been pushed out from this year to 2027 or further. Copper foil (HVLP) continues to be a

severe bottleneck, massively benefitting the few players capable of production – **Mitsui Kinzoku** (**5706 JP**) and **Co-Tech** (**8358 TT**) most importantly. Copper-clad laminate lead times have stretched even further, with Korean CCL import prices up 74.5% YoY as of last month, meaning **Elite Material Co** (**2383 TT**), **ITEQ** (**6213 TT**) are still likely to see earnings higher than estimates for even longer than consensus currently expects.

我们面临的所有严重瓶颈都在进一步恶化。T-Glass（T 级玻璃纤维）产能缓解的时间点已从今年推迟到 2027 年或更久。铜箔（HVLP）仍然是一个严重的瓶颈，这让少数具备生产能力的厂商大为受益——其中最重要的是三井金属（Mitsui Kinzoku，5706 JP）和台光电（Co-Tech，8358 TT）。覆铜板（CCL）的交货周期进一步拉长，截至上个月，韩国 CCL 进口价格同比上涨了 74.5%，这意味着台燿科技（Elite Material Co，2383 TT）和联茂电子（ITEQ，6213 TT）的盈利表现超过预期的时间可能比目前市场共识预期的还要长。

However, in our 26 Trades, we also said we'd continue monitoring areas of emerging interest, as well as ongoing bottlenecks.

然而，在我们的"26 个交易机会"中，我们也提到将继续监测新兴关注领域以及持续存在的瓶颈。

We believe there are two primary areas in the midstream and downstream that will continue to outperform: PCBs and IC Substrates. The reasoning is simply that both of these areas need to constantly be upgraded to keep up with the rising complexity necessary to scale AI.

我们认为，中游和下游有两个主要领域将继续表现优异：PCB（印制电路板）和 IC 载板。原因很简单，这两个领域都需要不断升级，以跟上扩展 AI 规模所必需的日益增长的复杂性。

PCBs and the Interconnect Bottleneck

PCB 与互连瓶颈

Nvidia is marketing the Rubin series as a platform consisting of Vera CPU, Rubin GPU, NVLink 6, ConnectX-9, BlueField-4, Groq LPU, and Spectrum-6.

Nvidia 正在将 Rubin 系列作为一个平台进行推广，该平台由 Vera CPU、Rubin GPU、NVLink 6、ConnectX-9、BlueField-4、Groq LPU 和 Spectrum-6 组成。

Over the last few generations, accelerator FLOPS have increased by roughly 300x, while interconnect performance has improved by closer to 10x. That gap is becoming the bottleneck. A GPU won't perform at its max capacity if data, signals, and power don't move between HBM, CPUs, NICs, DPUs, retimers, switch ASICs, optical modules, power delivery, and eventually other racks at a fast pace. This is why the PCB (printed circuit board) is being pulled from the background into the center of the AI supply chain. At rack scale, the board becomes part of the high-speed electrical channel.

在过去的几代产品中，加速器的算力（FLOPS）提升了约 300 倍，而互连性能的提升仅接近 10 倍。这一差距正成为瓶颈。如果数据、信号和电力不能在 HBM、CPU、NIC、DPU、重定时器（retimer）、交换机 ASIC、光模块、供电系统以及最终的其他机架之间快速流动，GPU 就无法发挥其最大效能。这就是为什么 PCB（印制电路板）正从幕后走向 AI 供应链的中心。在机架规模下，电路板已成为高速电气通道的重要组成部分。

For an accelerator to run at full utilization, the package and board stack have to scale with the silicon. This translates to larger substrates and PCBs with denser build-up layers, finer lines, more microvias, lower-loss materials, smoother copper, better glass cloth, and tighter warpage control.

为了让加速器以全利用率运行，封装和电路板堆栈必须随芯片同步扩展。这意味着需要更大尺寸的基板和 PCB，并具备更密的积层、更精细的线路、更多的微孔、更低损耗的材料、更光滑的铜箔、更优质的玻璃布以及更严格的翘曲控制。

We have spoken about this in **Interconnects 101**, back in July 2024, but the interesting second order impact that's happening due to the complexity of Interconnects is the stress on the printed circuit board.

我们在 2024 年 7 月的《互连基础 101》中讨论过这一点，但由于互连复杂性而产生的有趣的二阶效应，是其对印刷电路板造成的压力。

The Interconnect Layer

Interconnect layer	What it connects	Why it stresses PCBs/substrates
In-package interconnect	GPU/ASIC dies, HBM stacks, chiplets	Larger package substrates/interposers; fine-pitch RDL/build-up routing; dense bump/microvia escape; SAP/mSAP or other fine-line processes; warpage and CTE control
Board-level interconnect	GPU/OAM/SXM modules, CPU host boards, NICs/DPUs, retimers, power delivery	High-layer low-loss PCBs; controlled impedance; low-loss CCL; low-profile copper; via-stub/backdrill control; dense BGA/OAM escape
Rack/scale-up interconnect	Compute trays, GPU trays, NVLink/scale-up switch trays; sometimes midplanes/backplanes/ cabled backplanes	Longer high-speed channels; tight impedance/skew/crosstalk; very low insertion loss; high-speed connectors/cables; sometimes high-layer backplane PCBs
Network/scale-out interconnect	Switch ASICs, OSFP/QSFP-DD optical modules, DAC/ACC/AEC cables, NICs/DPUs	Switch-board and module-board loss budget; chip-to-module channel design; thermal/mechanical density at front panel; high-speed PCB materials and connector/cage design

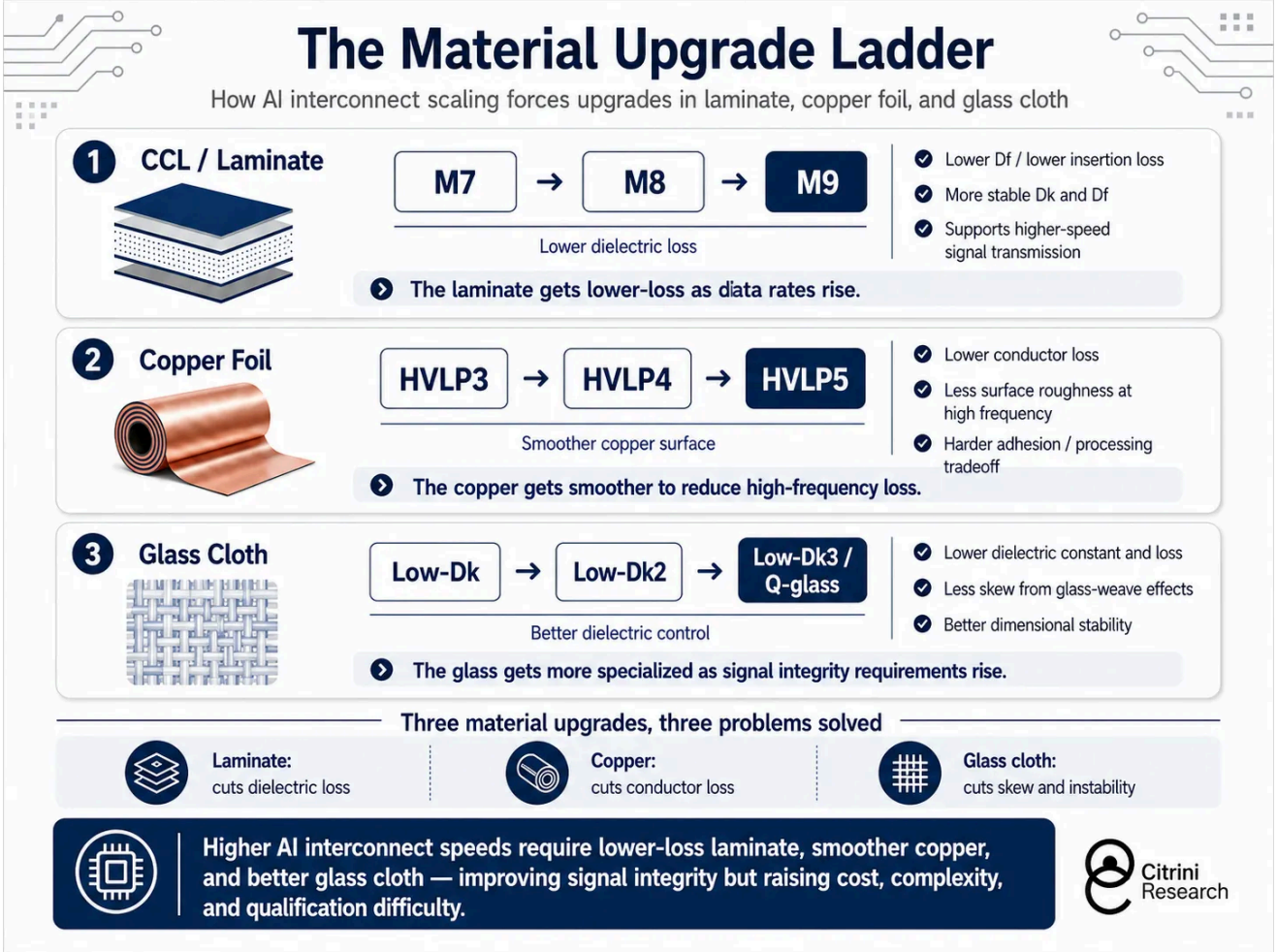
Citrini Research

As speed rises, the PCB material itself starts affecting performance. If the dielectric loss is too high, the signal attenuates. If copper roughness is too high, conductor loss rises. If the glass weave is uneven, the signal sees different dielectric environments and signal mismatch increases. This is why the PCB requires constant upgrades.

随着速率的提升，PCB 材料本身开始影响性能。如果介质损耗过高，信号就会衰减。如果铜箔粗糙度过高，导体损耗就会上升。如果玻纤布编织不均匀，信号会处于不同的介质环境中，导致信号失配增加。这就是 PCB 需要不断升级的原因。

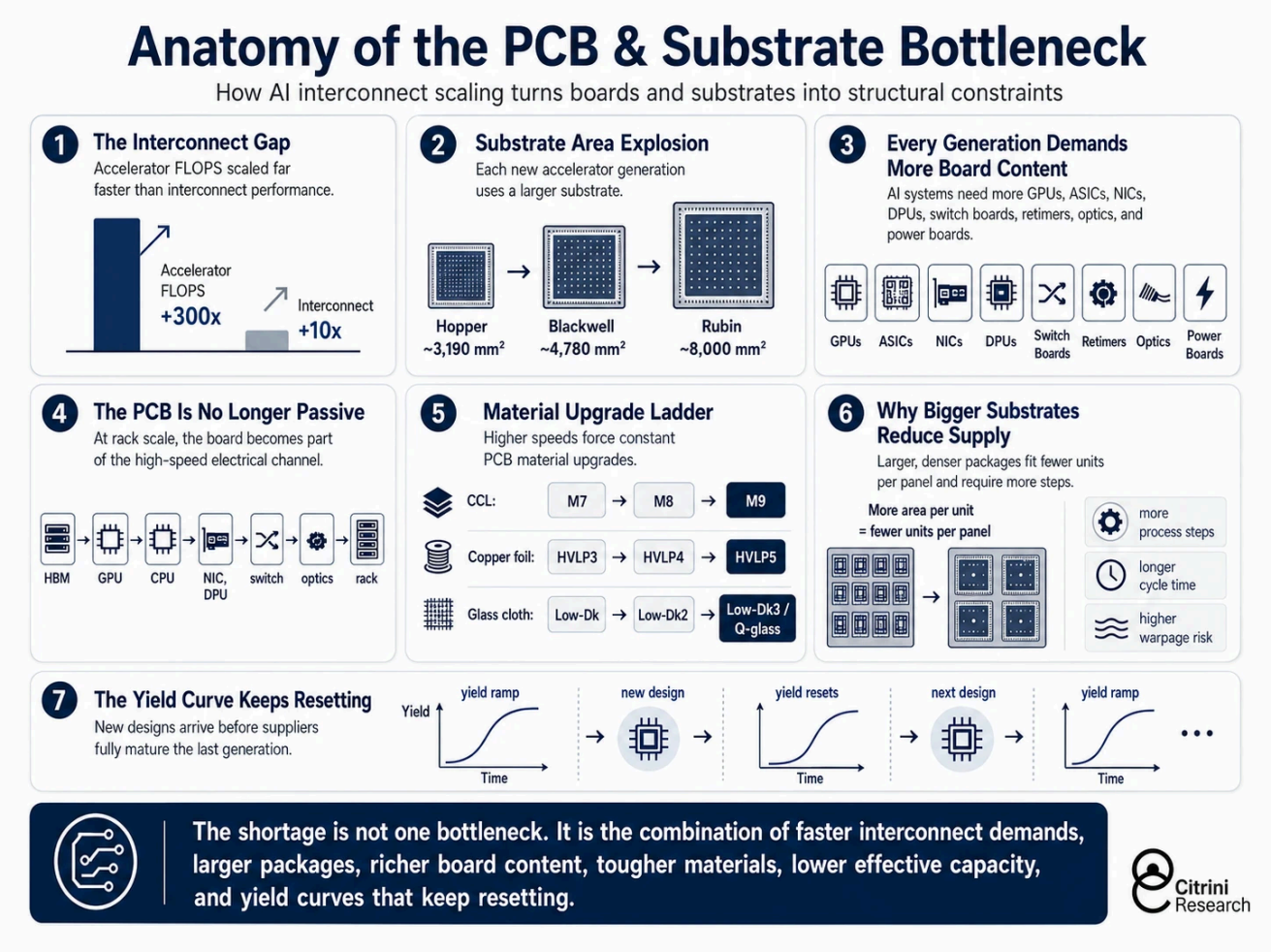
The CCL is upgrading from M7 to M8 to M9. The glass cloth is upgraded from Low-Dk to Low-Dk2 to Low-Dk3 or Q-glass. The copper foil in the CCL is upgraded from HVLP3 to HVLP4 to HVLP5. PCBs are transitioning to higher-layer HDI and selective mSAP/SAP-style fine-line processing and smaller ABF substrates are moving to larger ABF substrates and more SAP-intensive substrates.

CCL 正在从 M7 升级到 M8 再到 M9。玻璃布从 Low-Dk 升级到 Low-Dk2，再到 Low-Dk3 或 Q-glass。CCL 中的铜箔正在从 HVLP3 升级到 HVLP4 再到 HVLP5。PCB 正在向更高层数的 HDI 和选择性 mSAP/SAP 式精细线路工艺过渡，而较小的 ABF 载板正在向更大的 ABF 载板和更多采用 SAP 工艺的载板演进。



Every new AI accelerator asks the substrate to do more. More compute dies, more HBM, more chiplets, more I/O, more power, and more signals moving across a larger package.

每一款新型 AI 加速器都对基板提出了更高的要求。更多的计算芯片、更多的 HBM、更多的芯粒、更多的 I/O、更高的功耗，以及在更大的封装上进行传输的更多信号。



Nvidia GPU substrate area moved from roughly 3,190 mm^2 in Hopper to 4,780 mm^2 in Blackwell, and is expected to reach around 8,000 mm^2 in Rubin.

> Nvidia GPU 基板面积从 Hopper 的约 3,190 mm^2 增加到 Blackwell 的 4,780 mm^2，预计在 Rubin 中将达到约 8,000 mm^2。

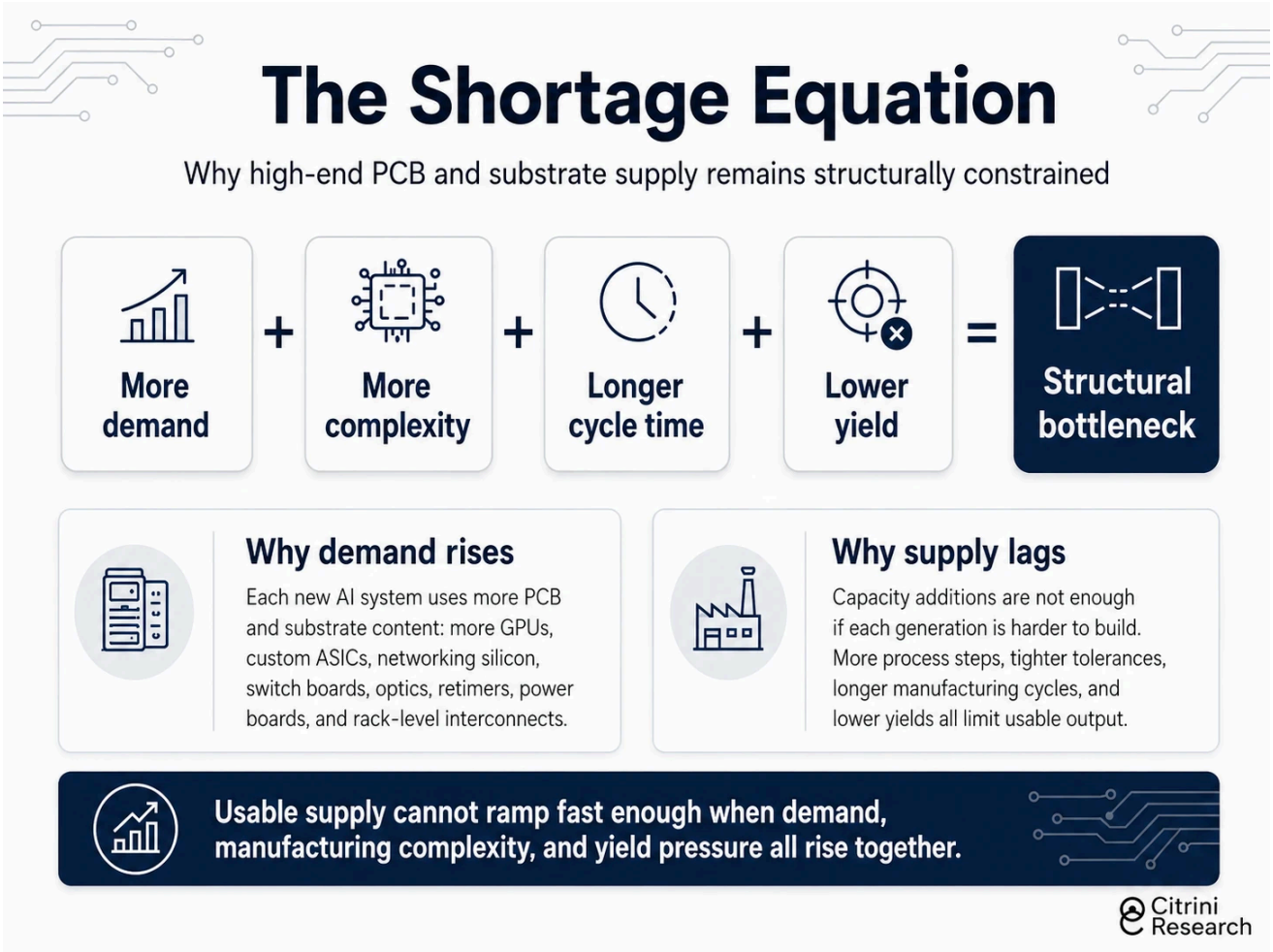
A bigger substrate takes more panel area, so the same factory produces fewer units. A more complex substrate takes more process steps, so the cycle time gets longer. A larger body is harder to keep flat, so warpage risk rises. More I/O means more routing, finer lines, more laser vias, more plating, and more SAP capacity. The demand goes up each generation because hyperscalers want more accelerators. The complexity goes up because each accelerator needs a larger, denser, more advanced substrate. The supply is getting squeezed from two sides. The unit demand is rising while the cycle time is

also increasing.

更大的基板占用更多的面板面积，因此同一工厂产出的单位数量会减少。更复杂的基板需要更多的工艺步骤，因此周期时间会变长。更大的主体更难保持平整，因此翘曲风险增加。更多的 I/O 意味着更多的布线、更细的线路、更多的激光通孔、更多的电镀以及更多的 SAP 产能需求。由于超大规模云服务商需要更多的加速器，需求每一代都在增长。由于每个加速器都需要更大、更密集、更先进的基板，复杂性也在不断提升。供应正受到双重挤压：单位需求在上升，而生产周期也在延长。

The same shift is happening at the board level. AI architecture is moving from the **server** as the basic unit to the **rack** as the basic unit. High-speed links increasingly run through switch trays, midplanes, backplanes, NIC and DPU boards, retimer boards, optics boards, and active copper assemblies. **This is forcing the material upgrade in PCBs.**

同样的转变也正在电路板层面发生。AI 架构正从以服务器为基本单元转向以机架为基本单元。高速链路越来越多地穿过交换机托盘、中板、背板、NIC 和 DPU 板、重定时器板、光学板以及主动铜缆组件。这正迫使 PCB 进行材料升级。



The shortage in high-end PCBs and substrates is due to a stack of bottlenecks. With every new accelerator generation, AI systems require more board and substrate content: more GPUs, custom ASICs, networking silicon, switch boards, optics, retimers, power boards, and rack-level interconnects. Demand is increasing not only because more AI systems are being built, but also because each system now carries more PCB and substrate content than the last generation.

高端 PCB 和基板的短缺是由一系列瓶颈叠加造成的。随着每一代新型加速器的推出，AI 系统对电路板和基板含量的需求都在增加：包括更多的 GPU、定制 ASIC、网络芯片、交换板、光学器件、重定时器、电源板以及机架级互连。需求的增长不仅是因为正在建造更多的 AI 系统，还因为现在每个系统所包含的 PCB 和基板含量都比上一代更高。

PCB Fabricators PCB 制造商



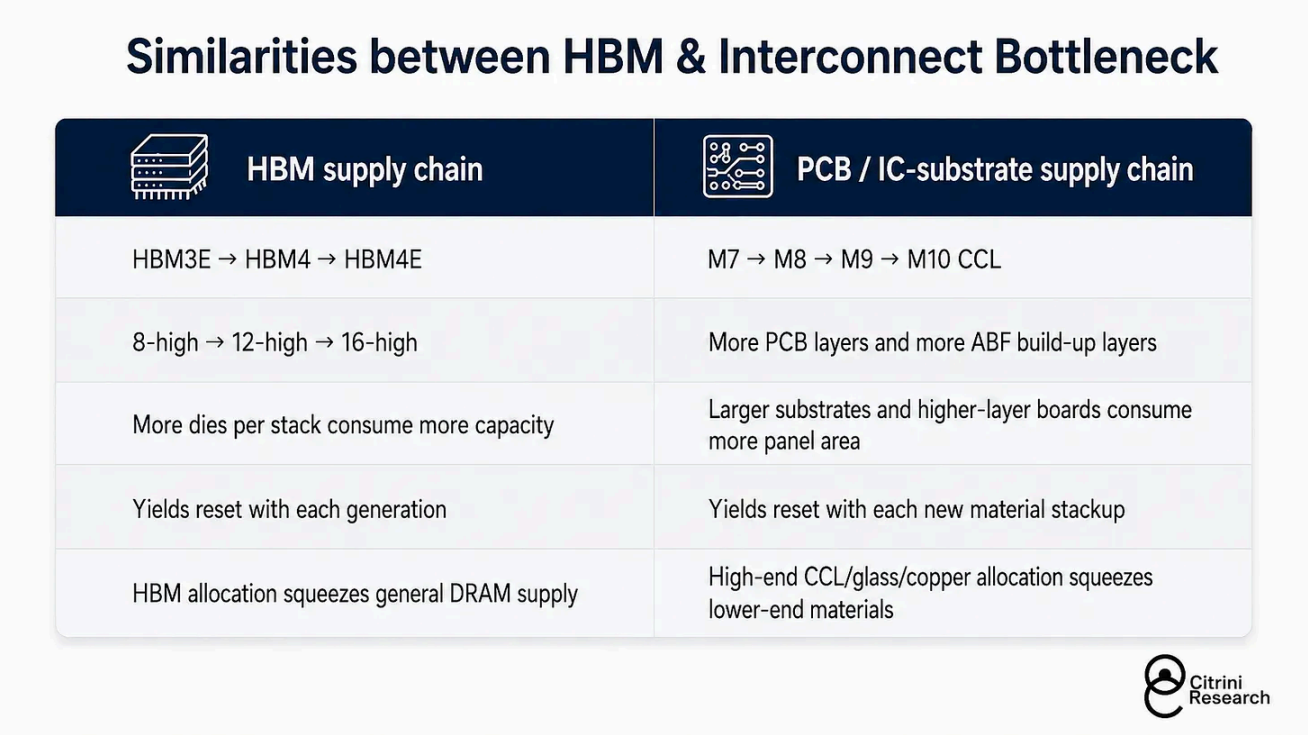
At the same time, each unit is becoming more complex and more time-consuming to manufacture. Substrates are getting larger, PCB layer counts are increasing, tolerances are becoming tighter, and more advanced materials are being used. This means capacity does not scale cleanly with demand. Even if suppliers add capacity, the number of process steps increases, the manufacturing cycles get longer, and the yields

go down.

与此同时，每个单元的制造正变得更加复杂且耗时。基板尺寸在变大，PCB 层数在增加，公差要求愈发严格，并且正在使用更先进的材料。这意味着产能无法随需求同步线性扩展。即便供应商增加产能，工艺步骤的增多、制造周期的延长以及良率的下降依然存在。

This is similar to what happened with HBM. Demand for HBM increased with every accelerator generation, but the product also became harder to manufacture. Each generation brought more dies per stack, more difficult packaging, and longer testing times. As complexity increased, yields became harder to maintain, especially during the transition from one generation to the next.

这与 HBM 领域发生的情况类似。随着每一代加速器的更迭，对 HBM 的需求不断增加，但产品的制造难度也随之攀升。每一代产品都意味着每个堆栈包含更多裸片、更复杂的封装工艺以及更长的测试时间。随着复杂性的提升，良率变得愈发难以维持，尤其是在从一代向下一代过渡的期间。



PCBs and substrates are now going through the same dynamic. Suppliers are being pushed to support new accelerator designs on a 12-15 month cadence, which gives them very little time to climb the yield curve before the next generation arrives. As a result, the yield curve keeps resetting. Demand rises, complexity rises, and usable

supply cannot ramp fast enough. That is what creates the structural bottleneck. We have covered the majority of the companies in AI Materials #5 in 26 Trades for 2026, but the companies below have the highest upside in our opinion.

PCB 和基板目前正经历着同样的动态。供应商被要求以 12-15 个月的节奏支持新的加速器设计，这使得他们在下一代产品到来之前，几乎没有时间提升良率曲线。结果就是，良率曲线不断重置。需求在增长，复杂度在提升，而可用供应的增长速度却无法跟上。这正是造成结构性瓶颈的原因。我们在《2026 年的 26 个交易机会》中的“AI 材料篇 #5”里已经涵盖了大部分公司，但我们认为以下公司具有最高的上涨空间。

IC Substrates IC 载板

Our top single pick for continued outperformance in AI Materials is **Ibiden** (**4062 JP**).

我们在 AI 材料领域持续看好的首选标的是 Ibiden (4062 JP)。

At the accelerator package level, the bottleneck begins at the IC substrate. AI GPUs and custom ASICs are getting physically larger because they need to connect more compute dies (2 dies or 4 dies), HBM stacks (6 to 8 to 12 stacks), chiplets, I/O dies, power-delivery structures, and the server board. As package size grows, each substrate consumes more ABF material, copper, T-glass, plating chemistry, and panel area. A production panel that previously yielded many smaller substrates now yields fewer large AI substrates. Even if factory capacity is unchanged, effective unit output falls.

在加速器封装层面，瓶颈始于 IC 载板。由于需要连接更多的计算裸片（2 片或 4 片）、HBM 堆栈（6、8 或 12 堆栈）、芯粒（chiplets）、I/O 裸片、供电结构以及服务器主板，AI GPU 和定制 ASIC 的物理尺寸正变得越来越大。随着封装尺寸的增加，每块载板会消耗更多的 ABF 材料、铜、T 玻璃、电镀化学品和面板面积。过去能产出许多小型载板的生产面板，现在产出的超大 AI 载板数量却变少了。即使工厂产能保持不变，实际的单位产出也会下降。

Emerging Upstream Bottleneck: Specialty Gases (WF_6)

新兴上游瓶颈：特种气体（WF_6）

Back in December, we highlighted the parabolic rise in Tungsten as an area to monitor for emerging bottlenecks in 2026.

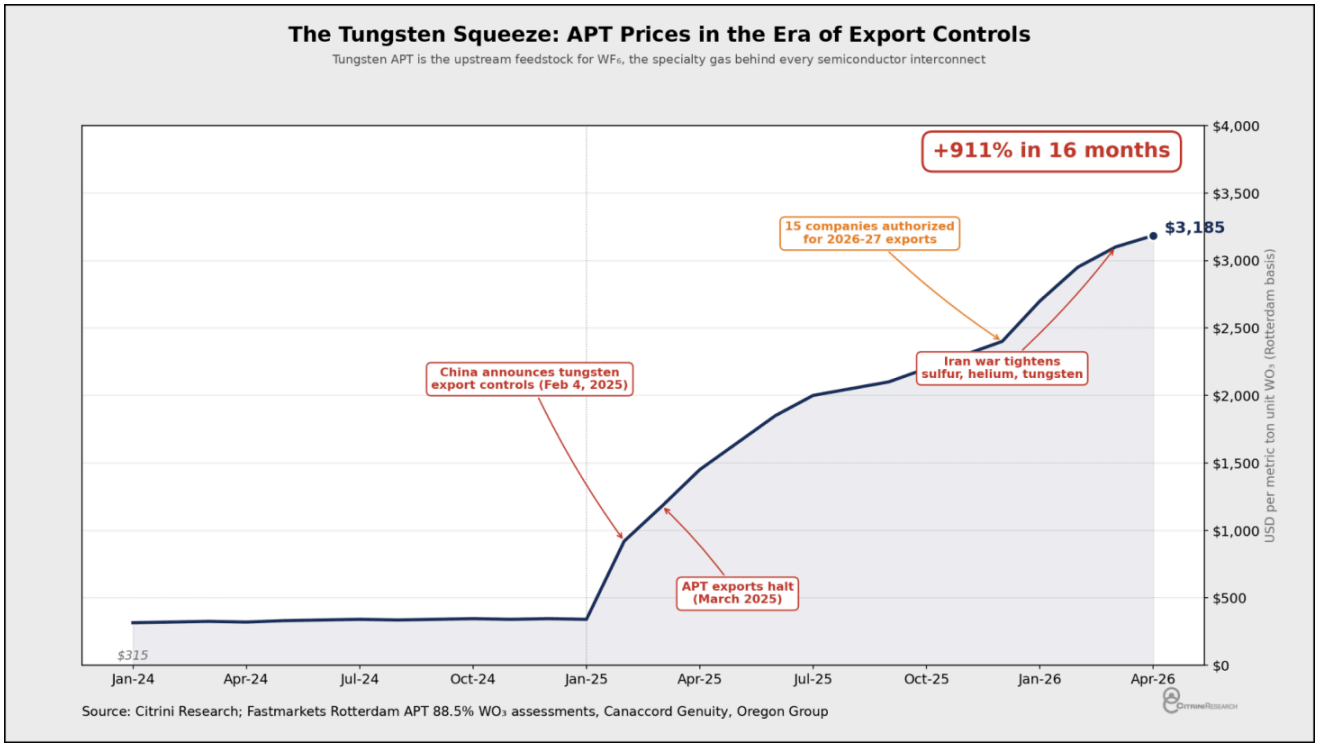
早在 12 月，我们就强调了钨（Tungsten）价格的抛物线式增长，并将其列为 2026 年需重点监测的潜在瓶颈领域。

Tungsten Hexafluoride (WF_6) is used in tungsten CVD for via fill and contact metallization at advanced nodes, and demand intensity scales directly with HBM stack height and 3D NAND layer count. Supply is concentrated across Kanto Denka, SK Materials (held by SK Inc, **034730 KS**), Merck KGaA (Formerly Versum), and Air Liquide, with the China localization push absorbing a meaningful share of domestic capacity. Contract pricing has been ratcheting higher for six straight quarters and the latest signals suggest the curve is steepening, not flattening, as we approach the HBM4 capacity ramp.

六氟化钨（WF_6）在先进节点的钨化学气相沉积（CVD）中用于通孔填充和接触金属化，其需求强度与 HBM 堆叠高度及 3D NAND 层数直接成正比。供应端主要集中在关东电化（Kanto Denka）、SK Materials（由 SK Inc, 034730 KS 持有）、默克集团（Merck KGaA，原 Versum）和液化空气（Air Liquide），而中国的本土化进程吸收了其国内产能的很大一部分。合约价格已连续六个季度攀升，最新信号表明，随着 HBM4 产能扩产周期的临近，价格曲线正在变得更加陡峭，而非趋于平缓。

Following supply chain disruptions due to the conflict in Iran, we're confident that we'll see this bottleneck become even more pronounced than it already is.

由于伊朗冲突导致的供应链中断，我们确信这一瓶颈将变得比现在更加严峻。



Kanto Denka (**4047 JP**) is the cleanest pure-play. Fluorine chemicals run roughly 35% of revenue but a much larger share of segment EBIT, and inside that segment the WF_6 and electronics gas business is the growth engine.

关东电化 (4047 JP) 是最纯粹的相关标的。氟化学品约占其营收的 35%，但在分部息税前利润（EBIT）中占比更高，而在该业务板块内部，WF_6 和电子气体业务是其增长引擎。

Korea Unlocked 韩国解锁

When we wrote about the coming ascendancy of the Korean stock market in August 2025's **State of the Themes**, there were limited ways that an international individual investor could express the trade – we contented ourselves with EWY. But now, Korean equities are available broadly. As we move forward, our Korea Semis analyst Jukan will be more freely examining Korean stocks (*especially because we won't have to deal with*

hundreds of people complaining they can't invest in them).

我们在 2024 年 8 月的《主题现状报告》（State of the Themes）中撰文讨论韩国股市即将崛起时，国际个人投资者参与该交易的途径还很有限——当时我们仅满足于通过 EWY 进行布局。但现在，韩国股票已获得广泛的准入渠道。展望未来，我们的韩国半导体分析师 Jukan 将更自由地研究韩国股票（尤其是因为我们不必再应付成百上千抱怨无法投资这些股票的人了）。

Within the bottleneck-hunting game, there is one area that is both the most obvious and still meaningfully under-discounted in the price — leading-edge foundry. AI accelerators, custom ASICs, HBM base dies, networking chips, CPUs, XPUs — all of them ultimately come off a leading-edge logic wafer. HBM, CoWoS, optics, and substrates have all become critical bottlenecks, but the starting point of every one of these bottlenecks is still leading-edge process capacity.

在寻找瓶颈的博弈中，有一个领域既是最显而易见的，但在价格中仍被显著低估——那就是先进制程代工。AI 加速器、定制 ASIC、HBM 基础底片（base dies）、网络芯片、CPU、XPU——所有这些最终都产自先进制程逻辑晶圆。尽管 HBM、CoWoS、光学器件和基板都已成为关键瓶颈，但每一个瓶颈的起点依然是先进制程的产能。

TSMC has already signaled that AI demand is exceptionally strong and that supply constraints persist on both leading-edge nodes and advanced packaging.

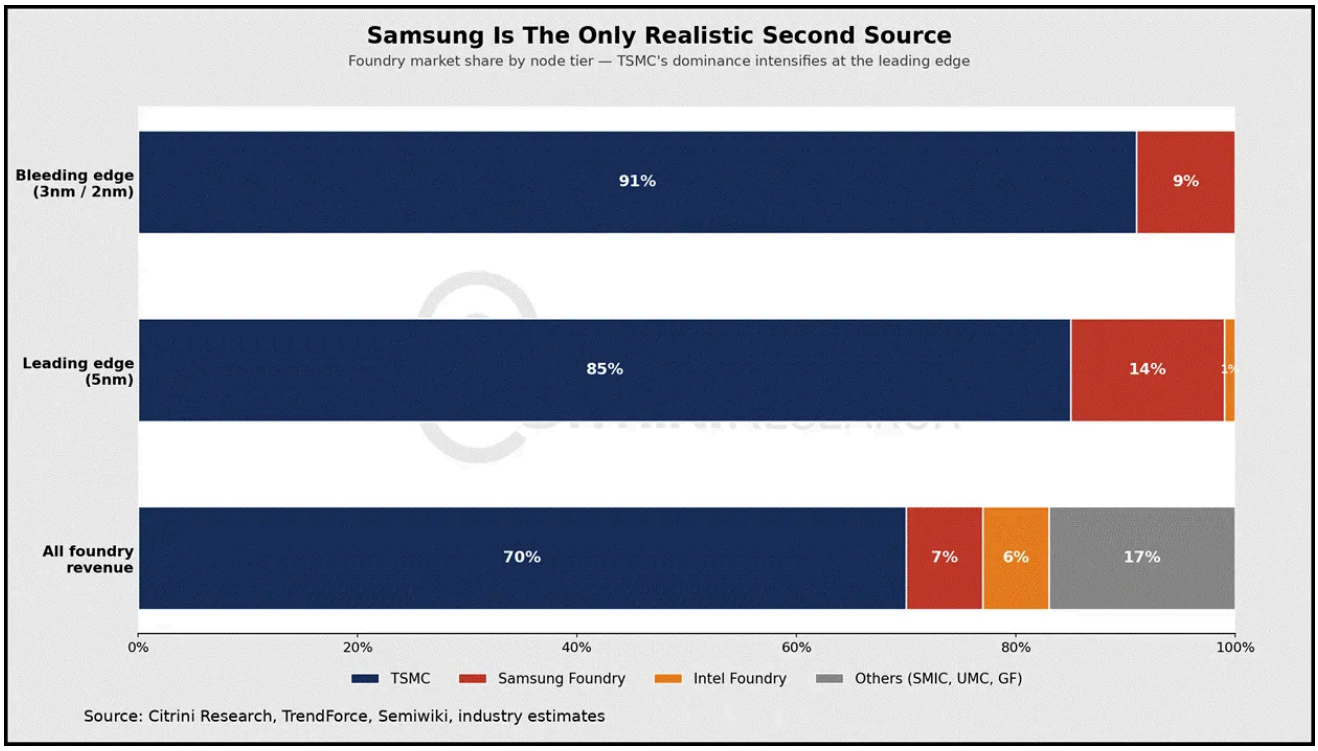
台积电已经发出信号，表明 AI 需求异常强劲，且领先制程节点和先进封装领域仍存在供应限制。

> *"Thus, we are using multiple levers to do everything we can, wherever we can, however we can to maximize the support to all our customers across all platforms.*
>
> *"因此，我们正利用多种手段，在任何可能的地方、以任何可能的方式竭尽所能，为所有平台上的所有客户提供最大程度的支持。"*
>
> *The demand are [sic] very robust, especially from the HPC and AI applications. And also, we try very hard to speed it up and pull in all the equipment as we can. Still our supply is*

> *very tight. Demand continues to increase.”*
>
> *需求非常强劲，特别是来自高性能计算（HPC）和人工智能（AI）应用的需求。同时，我们也在竭尽全力加快进度，并尽可能地调集所有设备。即便如此，我们的供应依然非常紧张。需求仍在持续增长。*
>
> *- C.C. Wei, Q1 2026 Earnings Call*
>
> *- 魏哲家，2026 年第一季度财报电话会议*

This is where the bull case on **Samsung Foundry** (**005930 KS**) begins. The Samsung Foundry thesis isn’t that Samsung beats TSMC over a short horizon but that **Samsung Foundry can be re-rated as the most realistic second source of leading-edge capacity.**

这就是三星代工（Samsung Foundry，005930 KS）看涨逻辑的起点。三星代工的核心论点并非三星能在短期内击败台积电，而是三星代工可以作为最现实的前沿产能第二供应源，从而获得估值的重新评级。



Intel Foundry’s existence cannot be ignored, of course. Intel holds powerful strategic assets — leading-edge manufacturing presence on US soil, advanced packaging, the x86 ecosystem, and government backing. From a geopolitical standpoint, Intel could become an extremely important player over the long run. For US customers in

particular, the strategic value of having a “domestic leading-edge foundry” option is undeniable.

当然，英特尔代工（Intel Foundry）的存在不容忽视。英特尔拥有强大的战略资产——位于美国本土的前沿制造能力、先进封装技术、x86 生态系统以及政府的支持。从地缘政治的角度来看，长远而言，英特尔可能成为一个极其重要的参与者。特别是对于美国客户来说，拥有一个“本土前沿代工”选项的战略价值是不可否认的。

But what customers want **right now** is not a long-term vision — it is an executable alternative that can actually tape out a product on time and carry it through to volume production. By that yardstick, Samsung is a more meaningful present-tense player than Intel. Samsung already has leading-edge foundry production experience, has been running real customer products on 4nm/3nm, and possesses a vertically integrated structure spanning memory, logic, and packaging. Intel Foundry, by contrast, still has to clear hurdles around expanding its external customer base, validating process reliability, and securing high-volume production references with marquee customers.

但客户目前需要的并非长远愿景，而是一个可执行的替代方案，能够真正确保产品按时完成流片并顺利进入量产。以此标准衡量，三星在当下是比 Intel 更有分量的参与者。三星已经拥有前沿的代工生产经验，其 4nm/3nm 工艺已在运行真实的客户产品，并拥有涵盖存储、逻辑芯片和封装的垂直整合结构。相比之下，Intel 代工服务（Intel Foundry）仍需克服诸多障碍，包括扩大外部客户群、验证工艺可靠性，以及获取顶级客户的大规模量产背书。

In short, Intel looks more like a long-dated call option on a US-centric foundry, while Samsung looks like the more realistic global second source capable of absorbing TSMC’s bottleneck *today*.

简而言之，Intel 更像是一个针对以美国为中心的代工厂的长期看涨期权，而 Samsung 则看起来更像是一个更现实的全球第二供应源，能够吸收 TSMC 如今面临的瓶颈。

The moment the market begins to acknowledge this possibility, Samsung Foundry stops being viewed as a structurally broken business unit and starts being repriced as a

scarce option within the AI infrastructure supply chain.

一旦市场开始承认这种可能性，三星代工（Samsung Foundry）将不再被视为一个结构性受损的业务部门，而是开始作为 AI 基础设施供应链中稀缺的期权被重新定价。

Per our industry checks, there is a meaningful probability that Samsung Foundry is involved in portions of the memory-adjacent custom silicon program tied to Marvell's recently rumored Google hyperscaler AI infrastructure work.

根据我们的行业调查，三星代工厂很有可能参与了与 Marvell 最近传闻的 Google 超大规模 AI 基础设施工作相关的部分存储相邻定制芯片项目。

Why does this project matter for Samsung Foundry? **As hyperscalers continue to expand the custom logic, memory expansion, CXL, and inference acceleration chips that surround the TPU, the number of programs competing for TSMC's leading-edge capacity keeps rising. In that environment, Samsung does not need to displace TSMC on flagship accelerators. It only needs to recover *enough* trust to absorb a portion of the AI/HPC peripheral logic, memory-adjacent ASICs, and second-source programs where schedule reliability and supply-chain diversification matter most.**

为什么这个项目对三星代工（Samsung Foundry）至关重要？随着超大规模云服务商不断扩展围绕 TPU 的定制逻辑、内存扩展、CXL 和推理加速芯片，竞争台积电（TSMC）前沿产能的项目数量持续攀升。在这种环境下，三星并不需要取代台积电在旗舰加速器上的地位。它只需要恢复足够的信任，以吸收一部分 AI/HPC 外围逻辑、内存相关 ASIC 以及对进度可靠性和供应链多元化要求最高的备选方案（second-source）项目。

So who stands to benefit from this project?

那么谁将从这个项目中受益呢?

Doosan Tesna (131970 KS)



Source: Citrini Research

Doosan Tesna is Korea's #1 system semiconductor wafer test OSAT, with the overwhelming majority of its revenue tied to 12-inch system semiconductor testing for Samsung Foundry. As the de facto default vendor for back-end testing of system semiconductors coming out of Samsung Foundry — Exynos, CIS, AP, MCU — Doosan Tesna's volumes scale directly with Samsung Foundry's utilization rate. Why now? Because the company is already moving. In October 2025, Doosan Tesna disclosed the acquisition of testing equipment worth KRW 171.4 billion.

Doosan Tesna 是韩国排名第一的系统半导体晶圆测试 OSAT（外包封测代工厂），其绝大部分收入都与三星代工厂（Samsung Foundry）的 12 英寸系统半导体测试挂钩。作为三星代工厂产出的系统半导体（包括 Exynos、CIS、AP、MCU）后端测试的事实上首选供应商，Doosan Tesna 的业务量直接随三星代工厂的产能利用率而波动。为什么要关注现在？因为该公司已经开始行动。2025 年 10 月，Doosan Tesna 披露了价值 1714 亿韩元的测试设备收购计划。

The stated rationale is Apple CIS. Samsung has broken into the Apple CIS supply chain that Sony had effectively monopolized, and the structure that has now settled in is straightforward — Samsung's Austin fab handles front-end wafers, Doosan Tesna handles wafer test, and Samsung handles packaging. The line goes live in 2H26. Mobile CIS is cyclical, but given the nature of Apple as a customer, lines that get qualified-in tend not to come out. This is a line that pulls Doosan Tesna's revenue base

up a notch.

所述理由是苹果的 CIS（图像传感器）。三星已成功打入此前由索尼几乎垄断的苹果 CIS 供应链，目前确立的结构非常明确——三星奥斯汀晶圆厂负责前道晶圆制造，Doosan Tesna 负责晶圆测试，三星负责封装。该产线将于 2026 年下半年投产。移动端 CIS 具有周期性，但鉴于苹果作为客户的特性，一旦通过验证的产线通常不会被撤换。这条产线将使 Doosan Tesna 的营收基数提升一个档次。

The problem is that the KRW 171.4 billion was just the start.

问题在于，1714 亿韩元仅仅是个开始。

On April 28, 2026, Doosan Tesna filed two separate disclosures. The first was a KRW 190.9 billion acquisition of Teradyne and Semes equipment. Teradyne ATE is for SoC testing, not memory. The destination, per our checks, is Nvidia's LPU. The chip leverages the inference acceleration tech Nvidia absorbed through its Groq "acquisition", with production nodes running on Samsung Foundry through the LP30/LP35 generations. Equipment installation wraps up by October 30, 2026, with meaningful revenue contribution kicking in from 4Q26. Applying the rule-of-thumb conversion rate the OSAT industry typically uses (~30% of capex translating to revenue) gets you to roughly KRW 50–60 billion of incremental annual revenue ($30–$40 million).

2026 年 4 月 28 日，Doosan Tesna 提交了两份独立的披露文件。第一份是关于斥资 1909 亿韩元收购 Teradyne 和 Semes 设备的公告。Teradyne ATE 用于 SoC 测试，而非内存测试。根据我们的核查，这些设备的去向是 Nvidia 的 LPU。该芯片利用了 Nvidia 通过"收购" Groq 所吸收的推理加速技术，生产节点在 Samsung Foundry 的 LP30/LP35 代工艺上运行。设备安装将于 2026 年 10 月 30 日完成，并从 2026 年第四季度开始产生实质性的营收贡献。按照 OSAT 行业通常使用的经验转换率（资本支出约 30% 转化为营收）计算，这将带来约 500 亿至 600 亿韩元的年度新增营收（约合 3000 万至 4000 万美元）。

The number itself isn't trivial, but what matters more is the nature of the customer. Until now, the market has bucketed Doosan Tesna as a Samsung Foundry-tied mobile AP/CIS testing house. The KRW 190.9 billion disclosure is the first crack in that frame. It's the moment a global tier-1 AI accelerator OEM enters Doosan Tesna's customer base directly — the moment its revenue mix starts shifting from a mobile cycle to an

AI infrastructure cycle.

这个数字本身并非微不足道，但更重要的是客户的性质。直到现在，市场一直将 Doosan Tesna 归类为与三星代工厂绑定的移动端 AP/CIS 测试厂商。这份 1909 亿韩元的披露公告是打破这一固有印象的第一道裂缝。这意味着全球一级 AI 加速器 OEM 厂商直接进入了 Doosan Tesna 的客户群——这一刻，其营收结构开始从移动周期向 AI 基础设施周期转型。

That said, the LPU bet has an expiration date. From the LP40 generation, Nvidia looks set to move production to TSMC. The revenue window the KRW 190.9 billion creates is powerful, but it isn't sticky.

即便如此，LPU 的押注是有期限的。从 LP40 代开始，Nvidia 似乎已准备好将生产转移到 TSMC。这 1909 亿韩元创造的营收窗口期虽然强劲，但并不具备粘性。

This is where Marvell's Google-bound memory-adjacent custom silicon — discussed earlier in this note — re-enters the picture. Marvell is a different animal from LPU. Set to be produced on Samsung's 4nm and slated to launch in mid-2027, this chip has a meaningful probability of migrating onto Samsung Foundry's 2nm in the next generation. That isn't a single-node window — it's a sticky workload that follows the node migration cycle. Marvell matters for Doosan Tesna not because it's "another large customer," but because it's the most plausible candidate to backfill the seat that LPU vacates.

这就是 Marvell 为谷歌定制的、与内存相关的定制芯片（本报告前文已讨论）再次登场的地方。Marvell 与 LPU 完全不同。该芯片定于 2027 年中期推出，采用三星 4nm 工艺生产，且极有可能在下一代迁移至三星代工厂的 2nm 工艺。这并非单一节点的生命周期，而是一个遵循节点迁移周期的粘性工作负载。Marvell 对 Doosan Tesna 而言至关重要，并非因为它仅仅是"另一个大客户"，而是因为它是填补 LPU 留下的空缺最合理的候选者。

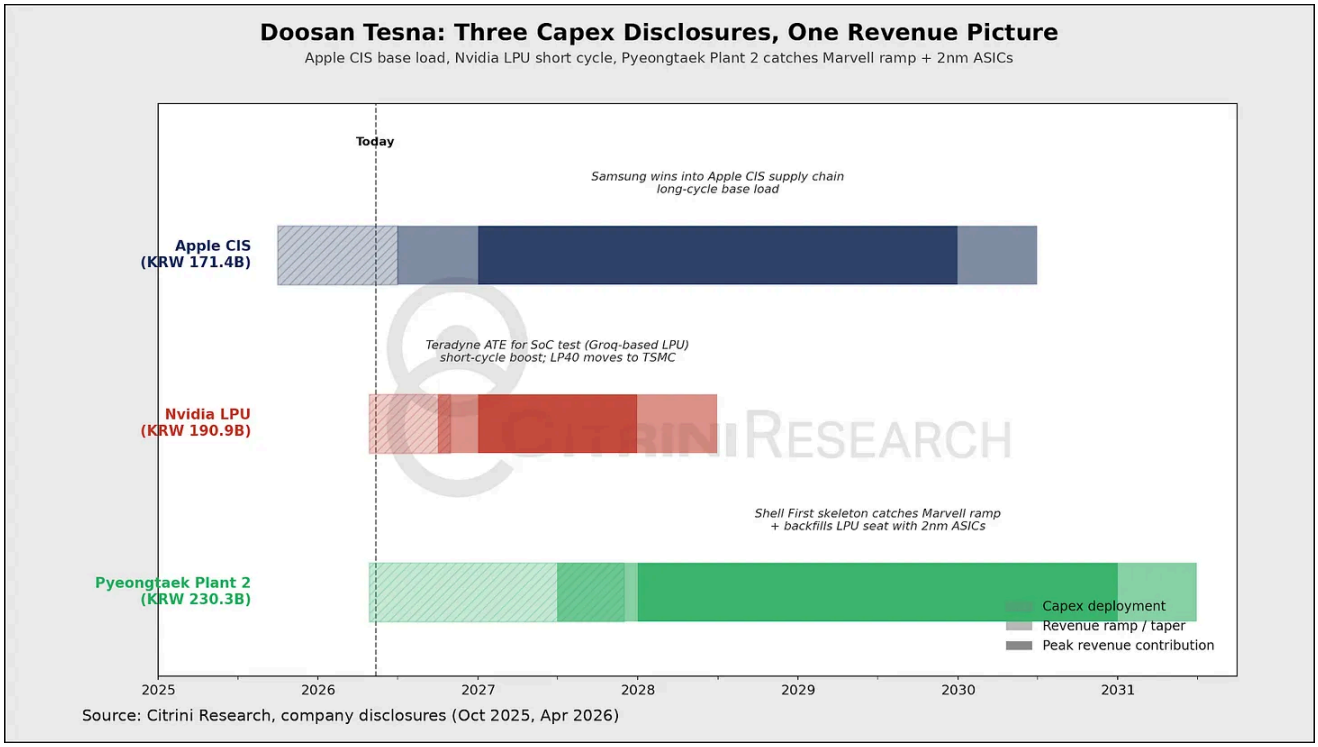
The company is already looking at this picture. The second disclosure on April 28 — the amendment to the Pyeongtaek Plant 2 capex disclosure — is the evidence. The investment amount was revised up from KRW 220.6 billion to KRW 230.3 billion, and the completion deadline was pushed from March 2027 to November 2027. Two points matter here. First, the construction that had been paused on softer 2025 demand has been resumed. Second, the project is being run on a Shell First framework — putting

up the building and clean room ahead of demand, with equipment capex deployed in stages as demand materializes.

公司已经在关注这一局面。4 月 28 日的第二次披露——即对平泽 2 号工厂资本支出披露的修订——就是明证。投资金额从 2206 亿韩元上调至 2303 亿韩元，完工期限从 2027 年 3 月推迟到 2027 年 11 月。这里有两个关键点：首先，此前因 2025 年需求疲软而暂停的建设已经恢复；其次，该项目正按照“外壳优先”（Shell First）的框架运行——即领先于需求建造厂房和无尘室，随着需求的实现再分阶段投入设备资本支出。

Shell First is the conservative skeleton OSAT players lay down when they have visibility on demand they're confident in. Restarting a paused project means the picture the company is now seeing for post-2028 back-end testing demand has shifted versus how it looked in 2025. And the November 2027 completion timing — landing just months after Marvell's mid-2027 launch, and overlapping with LP40's transition to TSMC — is what gives this disclosure its real meaning. The shell goes up in time to catch Marvell's ramp from day one and to receive whatever fills the void LPU leaves behind.

“外壳优先”（Shell First）是 OSAT 厂商在对需求有明确信心时所采取的保守骨架式布局。重启一个曾暂停的项目，意味着该公司目前对 2028 年后后端测试需求的预期，与 2025 年时的看法相比已发生了转变。而 2027 年 11 月的竣工时间点——恰好在 Marvell 2027 年年中发布后的几个月，并与 LP40 向 TSMC 的过渡期重叠——正是这一披露的真正意义所在。外壳工程的及时完工，既能从第一天起就赶上 Marvell 的产能爬坡，也能承接 LPU 撤出后留下的任何空白。



KRW 171.4 billion is Apple CIS coming online from 2H26 — base load. KRW 190.9 billion is LPU ramping hard from 4Q26 and tapering into LP40 — a short-cycle boost. Pyeongtaek Plant 2 is the long-cycle skeleton — sized to capture Marvell's ramp from its mid-2027 launch and to receive other 2nm ASICs into the seat LPU leaves behind. These are three stages of the same picture.

1,714 亿韩元代表苹果 CIS 业务从 2026 年下半年开始上线——这是基础负载。1,909 亿韩元则源于 LPU 从 2026 年第四季度开始的大幅放量，并逐渐过渡到 LP40——这是一个短周期的提振。平泽 2 号工厂则是长周期的骨架——其规模旨在承接 Marvell 从 2027 年中期启动的放量，并接收 LPU 退出后留下的其他 2nm ASIC 订单。这三个阶段共同构成了同一幅蓝图。

Doosan Tesna captures the back-end side of Samsung Foundry's recovery. But the foundry recovery does not start at wafer test. It starts earlier, at tool move-in.

Doosan Tesna 捕捉到了三星代工业务（Samsung Foundry）复苏带来的后端机遇。但代工厂的复苏并非始于晶圆测试，而是始于更早阶段的设备搬入。

Before Doosan Tesna tests Apple CIS, LPU, Marvell, or future 2nm ASICs, Samsung first has to install the front-end tools that make those wafers possible. That is where

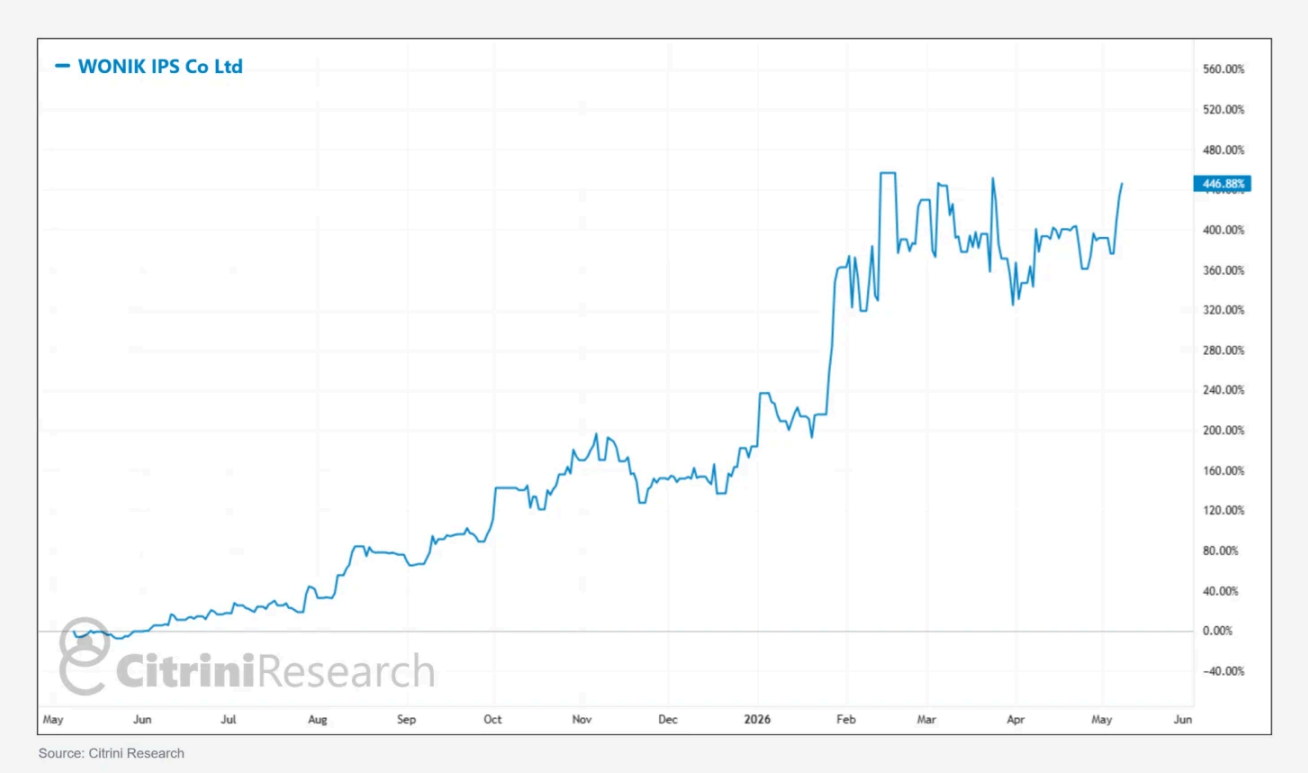
Wonik IPS enters the series.

在 Doosan Tesna 测试苹果的 CIS、LPU、Marvell 或未来的 2nm ASIC 之前，三星必须先安装能够生产这些晶圆的前端设备。这正是 Wonik IPS 在这一系列环节中登场的地方。

Wonik IPS (240810 KS)

This section benefited from research input and industry context from Damnag. https://substack.com/@damnang

本节内容得益于 Damnag 提供的研究投入和行业背景。https://substack.com/@damnang



Our second pick in the Samsung Foundry Beneficiary Series is Wonik IPS.

我们在三星代工受益者系列（Samsung Foundry Beneficiary Series）中挑选的第二个标的是 Wonik IPS。

If Doosan Tesna is the back-end testing beneficiary of Samsung Foundry's utilization recovery, Wonik IPS is the front-end equipment beneficiary of Samsung Foundry's

capacity build-out.

如果说 Doosan Tesna 是三星晶圆代工（Samsung Foundry）产能利用率恢复的后端测试受益者，那么 Wonik IPS 就是三星晶圆代工产能扩建的前端设备受益者。

Doosan Tesna gets paid when wafers start flowing. Wonik IPS gets paid before that happens. Before Tesla, Marvell, or any other 2nm ASIC customer can enter volume production, Samsung first has to fill Taylor with tools. EUV, etch, deposition, and metrology equipment go in before customer revenue appears. That timing difference is the core of the Wonik IPS thesis.

Doosan Tesna 在晶圆开始流片时获得报酬，而 Wonik IPS 在此之前就能获得报酬。在特斯拉、Marvell 或任何其他 2nm ASIC 客户进入量产之前，三星首先必须在泰勒工厂（Taylor）填满工具。在客户收入产生之前，极紫外光刻（EUV）、刻蚀、沉积和计量设备就已经入场了。这种时间差正是 Wonik IPS 投资逻辑的核心所在。

Wonik IPS matters because Samsung Foundry's leading-edge expansion is both a lithography and deposition story.

Wonik IPS 之所以重要，是因为三星晶圆代工厂的前沿扩张既涉及光刻技术，也涉及沉积技术。

Samsung's 3nm GAA mass production already used a new PECVD process based on Wonik IPS's GEMINI platform. EUV-compatible novel films, SADP films, and low-temperature hard-mask quality control were part of the process stack where Wonik IPS was cited. As Samsung moves from 3nm GAA to 2nm and expands Taylor around advanced logic customers, the process complexity does not fall. It rises. More GAA integration, more EUV-linked film requirements, more hard-mask control, and more multi-patterning complexity all increase the value of qualified deposition vendors already inside Samsung's advanced node ecosystem.

三星的 3nm GAA 量产已经采用了基于 Wonik IPS GEMINI 平台的新型 PECVD 工艺。EUV 兼容的新型薄膜、SADP 薄膜以及低温硬掩模质量控制，都是 Wonik IPS 被引用的工艺堆栈的一部分。随着三星从 3nm GAA 转向 2nm，并在泰勒工厂围绕先进逻辑客户进行扩张，工艺复杂度并未下降，反而有所上升。更多的 GAA 集成、更多与 EUV 相关的薄膜需求、更严格的硬掩模控制以及更复杂的多重曝光技术，都提升了已进入三星先进节点生态系统的合格沉积设备供应商的价值。

The relationship with Samsung runs deep. Samsung Electronics and Samsung Display each hold a 3.77% stake in Wonik IPS. Over 50% of total revenue is attributed to Samsung Electronics. Wonik IPS moves first when Samsung invests and goes quiet when Samsung pauses. The upside comes from this, and so does the risk.

与三星的关系源远流长。三星电子和三星显示器各持有 Wonik IPS 3.77% 的股份。其总营收的 50% 以上归功于三星电子。当三星进行投资时，Wonik IPS 率先行动；当三星暂停投资时，它也随之沉寂。增长空间源于此，风险亦然。

The reason to revisit Wonik IPS now is Taylor 1 becoming real, and Taylor 2 moving from a long-term option into an active review tied to 2nm customer discussions.

现在重新审视 Wonik IPS 的原因是泰勒 1 号厂（Taylor 1）正在变为现实，而泰勒 2 号厂（Taylor 2）也已从长期期权转变为与 2nm 客户洽谈挂钩的积极评估项目。

For a long time, Taylor was a fab without a customer. Samsung could build the shell, but without an anchor customer, equipment deliveries could be delayed. That was the problem.

长期以来，泰勒工厂一直处于“有厂无客”的状态。三星可以建成厂房外壳，但如果没有锚定客户，设备交付就可能会推迟。这就是问题的症结所在。

Tesla changed Taylor 1. The $16.5 billion long-term supply agreement turned Taylor from an empty strategic asset into a fab with an anchor workload. Samsung now expects Tesla-related chip production to begin in 2H27, with the fab entering its operational launch phase in 2026.

Tesla 改变了 Taylor 1。这项价值 165 亿美元的长期供应协议，将 Taylor 从一个空置的战略资产转变为一个拥有锚定工作负载的晶圆厂。三星目前预计与 Tesla 相关的芯片生产将于 2027 年下半年开始，该晶圆厂将于 2026 年进入运营启动阶段。

Taylor 2 is the next layer. Samsung has now begun an early review of a second Taylor fab, with the review being conducted in parallel with global customer order discussions for advanced 2nm logic. That matters because a second Taylor fab would not be a memory fab. It would be another leading-edge logic capacity addition. And

leading-edge logic capacity additions are exactly where deposition intensity rises.

Taylor 2 是下一个层面。三星目前已开始对泰勒市（Taylor）的第二座晶圆厂进行早期审查，该审查正与针对先进 2nm 逻辑芯片的全球客户订单洽谈同步进行。这一点至关重要，因为第二座泰勒晶圆厂将不会是存储芯片厂，而是另一个前沿逻辑芯片产能的增量。而前沿逻辑芯片产能的增加，恰恰是沉积工艺强度提升的地方。

Memory is not the core reason to own Wonik IPS, but it adds a second layer of upside.

存储业务并非持有 Wonik IPS 的核心理由，但它增加了第二层上行空间。

The foundry thesis alone is enough to put Wonik IPS on the map. But the setup is stronger because the company is not a pure foundry supplier. It also has **exposure to Samsung Memory and SK hynix capex**. If HBM-driven wafer allocation tightens conventional DRAM supply and forces another memory investment cycle through P4, P5, M15X, and Yongin, Wonik IPS gets a second lever.

仅凭代工业务的逻辑就足以让 Wonik IPS 备受关注。但由于该公司并非纯粹的代工供应商，其投资架构更为强劲。它还受益于三星存储（Samsung Memory）和 SK 海力士（SK hynix）的资本支出。如果 HBM 驱动的晶圆分配收紧了传统 DRAM 的供应，并迫使通过 P4、P5、M15X 和龙仁（Yongin）开启另一轮存储投资周期，Wonik IPS 将获得第二个增长杠杆。

Samsung Electro-Mechanics (009150 KS)

三星电机 (009150 KS)



Source: Citrini Research

If the Samsung Foundry thesis can be summarized in one line as “the most realistic call option on TSMC overflow,” then Samsung Electro-Mechanics is the lowest-strike option on that trade. Every time Samsung Foundry wins a new AI/HPC program, the chip must pass through two stages before it can ship — FC-BGA substrate and MLCC — and Samsung Electro-Mechanics holds first-tier vendor status in both.

如果说三星代工（Samsung Foundry）的投资逻辑可以被概括为“台积电产能溢出最现实的看涨期权”，那么三星电机（Samsung Electro-Mechanics）就是该交易中行权价最低的期权。每当三星代工赢得一个新的 AI/HPC 项目时，芯片在出货前必须经过两个阶段——FC-BGA 封装基板和 MLCC（多层陶瓷电容器）——而三星电机在这两个领域都拥有顶级供应商地位。

Once a leading-edge logic wafer leaves the fab, the next bottlenecks are, first, advanced packaging, and second, substrate (although the first is moot here, since most chips currently produced on Samsung Foundry’s leading-edge nodes do not use AVP).

一旦领先的逻辑晶圆离开晶圆厂，接下来的瓶颈首先是先进封装，其次是基板（尽管前者在这里并无实际意义，因为目前在三星代工厂领先节点上生产的大多数芯片并不使用先进封装）。

AI accelerators, custom ASICs, HBM base dies, networking chips — all of them must ultimately sit on a high layer-count FC-BGA. Global FC-BGA capacity has already

begun to tighten. Major suppliers — Ibiden, Unimicron, Nan Ya PCB, AT&S — are all prioritizing allocation toward AI substrates, and the ramp-up timeline for currently announced CAPEX is, at the earliest, mid-to-late 2027.

AI 加速器、定制 ASIC、HBM 基础晶圆、网络芯片——所有这些最终都必须承载在高层数的 FC-BGA 基板上。全球 FC-BGA 产能已经开始趋紧。主要供应商——Ibiden、欣兴电子（Unimicron）、南亚电路板（Nan Ya PCB）、奥特斯（AT&S）——都在优先将产能分配给 AI 基板，而目前已公布的资本支出（CAPEX）所带来的产能爬坡时间节点，最早也要到 2027 年中后期。

Within this structure, Samsung Electro-Mechanics' position is not merely that of a global top-tier supplier, but rather **the de facto captive first-tier substrate vendor for any chip taped out at Samsung Foundry**. In other words:

在此架构中，三星电机（Samsung Electro-Mechanics）的地位不仅是全球顶尖供应商，更是任何在三星代工厂（Samsung Foundry）流片的芯片事实上的专属一级基板供应商。换句话说：

● If the scenario in which Samsung Foundry participates in MRVL's Google hyperscaler memory-adjacent ASIC program materializes, the primary FC-BGA order for that chip is most likely to flow to Samsung Electro-Mechanics.

● 如果三星代工厂（Samsung Foundry）参与 MRVL 的 Google 超大规模内存相邻 ASIC 项目这一情景成为现实，那么该芯片的主要 FC-BGA 订单极有可能流向三星电机（Samsung Electro-Mechanics）。

● Every time Samsung Foundry secures additional external customer wins on 4nm/3nm/2nm GAA, new SKUs accumulate within the Samsung Electro-Mechanics FC-BGA mix.

● 每当三星代工厂（Samsung Foundry）在 4nm/3nm/2nm GAA 工艺上赢得更多外部客户时，三星电机（Samsung Electro-Mechanics）的 FC-BGA 产品组合中就会积累新的 SKU。

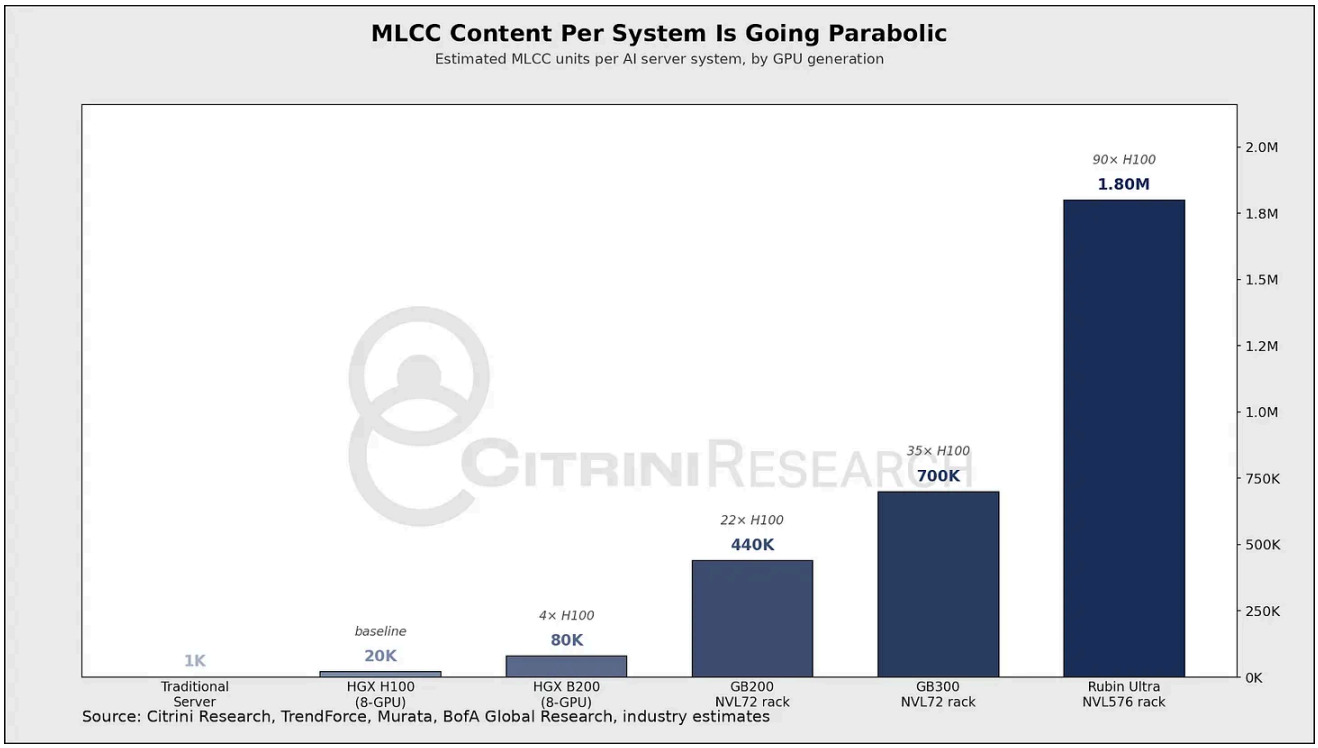
Put differently, the further Samsung Foundry normalizes, the more FC-BGA references Samsung Electro-Mechanics accumulates, laying the groundwork to penetrate the high-end segment that has so far been blocked off by the incumbents

(Ibiden, Unimicron).

换句话说，三星代工厂的运营越趋于常态化，三星电机积累的 FC-BGA 参考案例就越多，从而为其进入目前被老牌厂商（Ibiden、欣兴电子）所垄断的高端市场奠定基础。

Our preference for Samsung Electro-Mechanics is not driven solely by the Samsung Foundry normalization theme. If anything, **we believe the MLCC shortage cycle will have a larger impact on forward earnings**.

我们对三星电机的青睐并非仅仅受三星代工厂常态化这一主题驱动。如果说有什么不同的话，我们认为 MLCC 的短缺周期将对远期收益产生更大的影响。



The MLCC shortage cycle is a variable entirely independent of the Samsung Foundry thesis. But the fact that both variables converge on the same name is precisely what makes the risk/reward on Samsung Electro-Mechanics asymmetric.

MLCC 短缺周期是一个完全独立于三星代工论点的变量。但这两个变量汇聚在同一个名字上，正是三星电机（Samsung Electro-Mechanics）具备非对称风险回报比的原因。

MLCC content per AI server is increasing exponentially. Industry checks indicate that MLCC content per Rubin Ultra system is roughly 1.8 million units, approximately 90x the level of H100. MLCC lead times have already extended to 24 weeks — about 3x the actual production cycle of 6–8 weeks. This points to a tight supply environment in which strong downstream demand leaves little room to build inventory.

单台 AI 服务器的 MLCC 含量正在呈指数级增长。行业调查显示，每台 Rubin Ultra 系统的 MLCC 含量约为 180 万颗，约为 H100 水平的 90 倍。MLCC 的交货周期已延长至 24 周——约为 6 至 8 周实际生产周期的 3 倍。这表明供应环境趋于紧张，强劲的下游需求使得建立库存的空间微乎其微。

That said, the more important point in the Samsung Electro-Mechanics bull thesis goes one step further. Specifically, **if AI server power delivery transitions from 48V to 800V DC, an entirely new demand layer for ultra-high-voltage MLCCs is added on top**.

即便如此，三星电机（Samsung Electro-Mechanics）看涨逻辑中更重要的一点则更进了一步。具体而言，如果 AI 服务器的供电架构从 48V 转向 800V 直流电（DC），那么在现有需求之上，将增加一个全新的超高压 MLCC 需求层级。

Under the current 48V power architecture, MLCC demand can be broken into two broad buckets. The first is MLCCs used in power conversion and voltage stabilization. The second is MLCCs used around the major silicon — GPUs, ASICs, CPUs — for decoupling and noise suppression. The shortage that the market is currently focused on is largely concentrated in the latter bucket: high-reliability MLCCs supporting high-performance silicon.

在目前的 48V 供电架构下，MLCC 的需求大致可以分为两大类。第一类是用于功率转换和稳压的 MLCC。第二类是用于主要芯片（GPU、ASIC、CPU）周边，起去耦和噪声抑制作用的 MLCC。目前市场关注的短缺问题主要集中在后者：即支持高性能芯片的高可靠性 MLCC。

The story changes once power delivery transitions to 800V DC. A portion of the MLCCs that previously played the voltage stabilization role under 48V architecture will need to be replaced by ultra-high-voltage MLCCs capable of handling far higher voltages. In other words, the shift in AI server power architecture does not simply increase MLCC demand around the silicon — it **creates additional high-voltage**

MLCC demand at the power conversion stage as well.

一旦供电架构转向 800V 直流电，情况就会发生变化。在 48V 架构下原本承担稳压作用的一部分 MLCC，将需要更换为能够处理更高电压的超高压 MLCC。换句话说，AI 服务器供电架构的转变不仅增加了芯片周边的 MLCC 需求，还在功率转换阶段创造了额外的超高压 MLCC 需求。

This matters because 1,000V-class ultra-high-voltage MLCCs are meaningfully more expensive than standard 100V-class MLCCs. On a retail price basis, the cost gap is estimated to be greater than 10x.

这一点至关重要，因为 1,000V 级超高压 MLCC 的价格明显高于标准的 100V 级 MLCC。根据零售价格计算，两者的成本差距估计超过 10 倍。

For Samsung Electro-Mechanics, this is a highly significant structural shift. The company is already one of the major high-end MLCC suppliers globally alongside Murata, TDK, and Taiyo Yuden, with particular strength in high-reliability MLCCs for IT, automotive, and industrial applications. As a result, the expansion of demand for high-reliability and high-voltage MLCCs into AI servers feeds directly into Samsung Electro-Mechanics' ASP and margin structure.

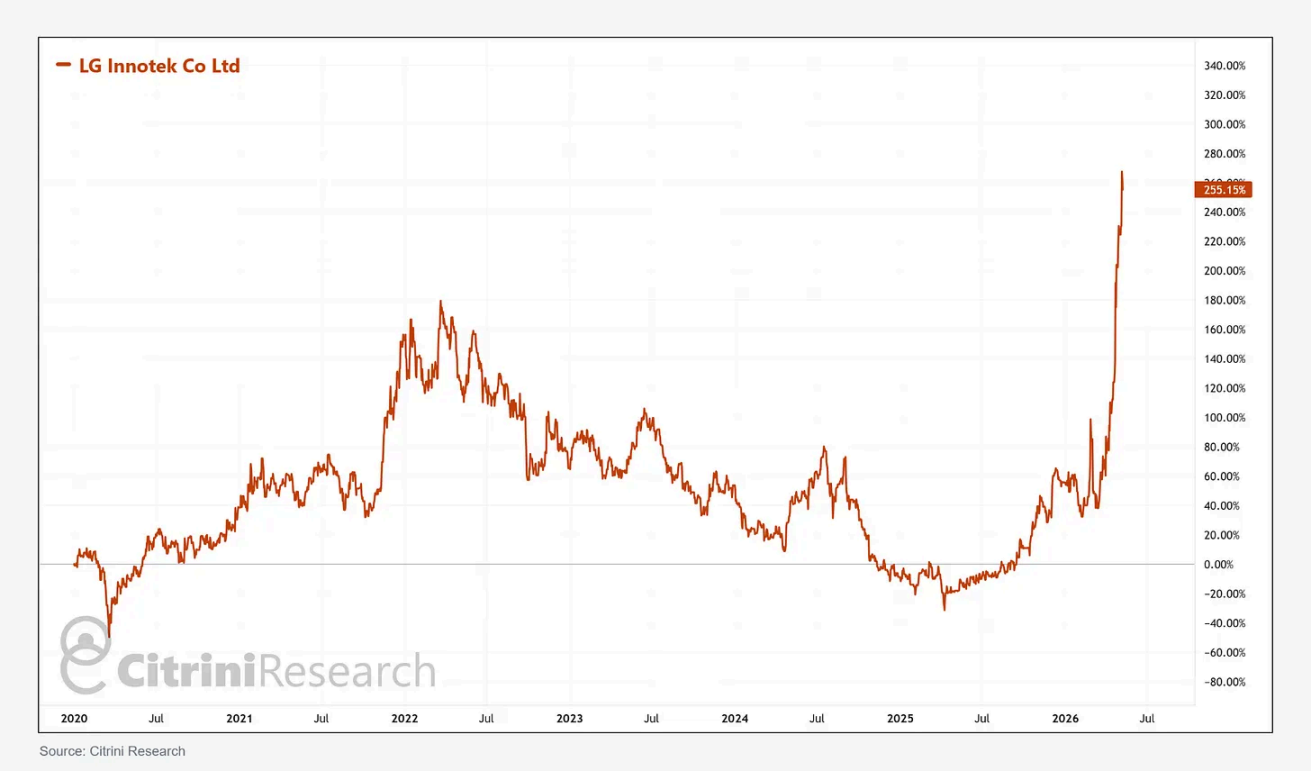
对于三星电机（Samsung Electro-Mechanics）而言，这是一次意义重大的结构性转变。该公司已与村田制作所（Murata）、TDK 和太阳诱电（Taiyo Yuden）并列为全球主要的高端 MLCC 供应商之一，在用于 IT、汽车和工业应用的高可靠性 MLCC 领域实力尤为雄厚。因此，AI 服务器对高可靠性和高电压 MLCC 需求的扩大，将直接提升三星电机的平均销售价格（ASP）和利润结构。

In particular, once the MLCC shortage — which began in high-reliability server-grade products — forces suppliers to reallocate capacity toward AI servers, the supply tightness is likely to spill over into adjacent end markets: mobile, automotive, and industrial. This is the same pattern we have already observed in both the memory and glass fiber cycles. For Samsung Electro-Mechanics, this generates two distinct effects: first, an increase in revenue from high-end AI server-grade MLCCs; and second,

improved pricing power across the entire MLCC portfolio.

特别是，一旦最初始于高可靠性服务器级产品的 MLCC 短缺迫使供应商将产能重新分配给 AI 服务器，供应紧张局势很可能会波及到相邻的终端市场：移动、汽车和工业领域。这与我们已经在存储器和玻璃纤维周期中观察到的模式如出一辙。对于三星电机（Samsung Electro-Mechanics）而言，这将产生两个明显的影响：首先，来自高端 AI 服务器级 MLCC 的营收将有所增加；其次，整个 MLCC 产品组合的定价能力将得到提升。

LG Innotek (011070 KS)



We close out the *Korea Unlocked* series with LG Innotek.

我们以 LG Innotek 为“解锁韩国”系列画上句号。

While LG Innotek's direct exposure to Samsung Foundry is limited, the name most compactly captures the structural reconfiguration unfolding across Korea's tech supply chain.

虽然 LG Innotek 对三星代工厂（Samsung Foundry）的直接敞口有限，但这个名字最简洁地体现了韩国科技供应链正在进行的结构性重组。

SiP Substrates: A Bottleneck Manufactured by Supplier Exit

SiP 基板：供应商退出制造的瓶颈

Every major substrate player globally is retooling capacity toward server- and AI-grade FC-BGA. What the market has yet to price in is the second-order consequence: as the entire industry migrates to FC-BGA, SiP substrates for smartphones, wearables, TWS, and AR/VR devices are entering structural undersupply.

全球各大基板厂商都在将其产能转向服务器级和 AI 级 FC-BGA。市场尚未定价的是其二阶后果：随着整个行业向 FC-BGA 迁移，用于智能手机、可穿戴设备、TWS 和 AR/VR 设备的 SiP 基板正进入结构性供应短缺。

Without the alphabet soup - big players chase high margin shiny object with all their resources, stop making lower margin older stuff, supply of older stuff (even though it's not seeing AI level demand) contracts, any sort of new demand or even just existing demand must be fulfilled, older stuff becomes higher margin for the few players left making it.

抛开那些复杂的术语不谈——大玩家们正倾尽所有资源追逐高利润的新兴产品，停止生产利润较低的旧型号，导致旧产品的供应收缩（尽管其需求并未达到 AI 级别）。然而，任何新增需求甚至仅仅是现有需求都必须得到满足，这使得对于少数仍在坚持生产的玩家来说，旧产品的利润反而变得更高了。

This is the same playbook we've already seen in DRAM and T-Glass.

这是我们在 DRAM 和 T-Glass 领域已经见过的同样套路。

In the Macronix case, the major NAND vendors concentrated on eSSD, TLC, and QLC while exiting legacy eMMC, SLC, and MLC NAND, leaving Macronix — the legacy NAND specialist — in de facto command of that market. In the Nanya Technology case, the DRAM Big 3 redirected capacity into DDR5, creating a DDR4 shortage severe

enough that NAND makers took equity stakes in Nanya.

在旺宏电子（Macronix）的案例中，主要的 NAND 厂商集中精力于 eSSD、TLC 和 QLC，同时退出了传统的 eMMC、SLC 和 MLC NAND 领域，这使得作为传统 NAND 专家的旺宏电子实际上掌控了该市场。在南亚科技（Nanya Technology）的案例中，DRAM 三巨头将产能转向 DDR5，导致 DDR4 严重短缺，以至于 NAND 制造商纷纷入股南亚科技。

The point is not that SiP TAM is growing at AI-like rates. On demand growth alone, this thesis looks thin. But when suppliers are actively exiting the category, modest demand growth is more than enough for survivor economics to play out. Currently, margins are projected to stay low for LG Innotek - around 9-9.5%. If we see the same kind of squeeze play out here as we did in MLC NAND, we think LG Innotek could see 2028 gross margins as high as 14% - currently expected to be 9.5%.

核心观点并非 SiP 的市场总量（TAM）正以类似 AI 的速度增长。仅凭需求增长，这一论点显得有些单薄。然而，当供应商正积极退出该领域时，适度的需求增长就足以让“幸存者经济”发挥作用。目前，LG Innotek 的利润率预计将维持在 9-9.5% 左右的低位。如果我们在这里看到与 MLC NAND 领域相同的挤压效应，我们认为 LG Innotek 到 2028 年的毛利率可能高达 14%——而目前的预期仅为 9.5%。

Updating Previous Semi Memo Ideas

更新之前的半导体备忘录观点

This is our third Semis Memo since we began the series on the thesis that the AI-related semiconductor industry would begin dominating the narrative.

这是我们自开始这一系列研究以来的第三份半导体备忘录，其核心论点是人工智能相关的半导体行业将开始主导市场叙事。

We are always on the lookout for the next big sub-theme for the AI infrastructure trade. We’ve done this since we first began our coverage of it with **Artificial**

Intelligence: Global Equity Beneficiaries.

我们一直在寻找人工智能基础设施交易中的下一个重大子主题。自我们发布《人工智能：全球股票受益者》并首次开始对该领域进行覆盖以来，我们就一直在这样做。

However, we've picked up a lot of new readers since then, so we figured it is a good time to remind everyone of something: just because we're on the lookout doesn't mean that anything is wrong with what we've already covered. Just like you could have bought the names we listed in our May 2023 AI primer and never bothered to figure out another bottleneck again (and roughly 6x'd from that), we want to make it clear that *it's our job to sniff out the next mispricing.* Nothing about that means that the previous mispricings have fully resolved. As our current dynamic AI allocation shows, we're still quite bullish on many of the ideas we've covered YTD.

然而，自那以后我们增加了很多新读者，因此我们认为现在是提醒大家一件事的好时机：仅仅因为我们在寻找新机会，并不意味着我们已经覆盖过的内容出了问题。就像你本可以购买我们在 2023 年 5 月的人工智能入门指南中列出的标的，并且再也不必费心去寻找另一个瓶颈（并从中获得大约 6 倍的收益）一样，我们想明确的是，发掘下一个定价错误是我们的职责。这并不意味着之前的定价错误已经完全修复。正如我们当前的动态人工智能配置所示，我们仍然非常看好今年以来所涵盖的许多投资想法。

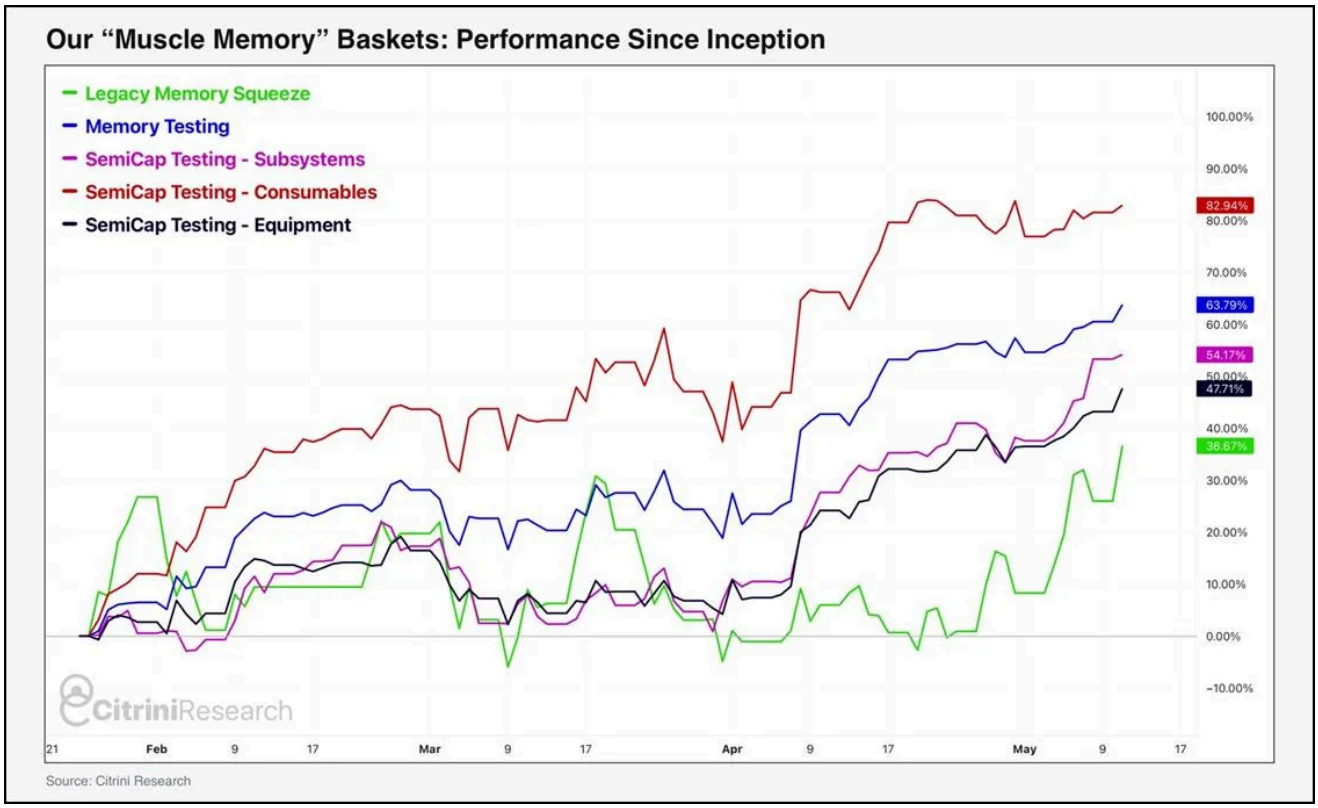
So let's look at how our previous collection of ideas from our first two Semis Memos have done – of course, you could have pretty much put all of SOXX on a dartboard, thrown a dart and pulled a pretty solid return buying whatever it landed on. This is less of a victory lap (not that we are averse to victory lapping, it's just in bad taste when it's this straightforward) and more focused on presenting what's performed better in relative terms to highlight where the opportunities still are.

让我们回顾一下前两篇半导体备忘录（Semis Memos）中收集的观点表现如何——当然，你几乎可以把整个 SOXX 指数放在飞镖盘上，随便投一镖，买入投中的任何标的都能获得相当可观的回报。这与其说是一场胜利巡礼（倒不是我们反感庆祝胜利，只是在行情如此直白的情况下这样做显得格调不高），倒不如说我们更专注于呈现哪些标的在相对收益上表现更好，以此凸显机会依然存在的领域。

Our January Memo was titled "**Muscle Memory**" and examined the best second order beneficiaries of the Memory Supercycle. Namely, these were Memory Testing, Semicap

Subsystems and the Legacy Memory Squeeze.

我们一月份的备忘录题为“肌肉记忆”，探讨了存储超级周期的最佳二阶受益者。具体而言，这些领域包括存储测试、半导体设备子系统以及传统存储供应收紧带来的机遇。



Our Memory Testing was led by **Aehr Test Systems** (**AEHR US**). And despite being up more than 300% since when we bought, we’re staying long. In fact, we’re adding.

我们的存储测试领域由 Aehr Test Systems (AEHR US) 领衔。尽管自我们买入以来其涨幅已超过 300%，但我们仍将继续持有多头头寸。事实上，我们正在加仓。

AEHR reported their lead hyperscale AI customer (GOOG) placed a $41 million follow-on production order for package-level burn-in of custom ASICs (TPUs), with deliveries starting in fiscal 2027. That puts second-half bookings above $92 million with six weeks still left in the quarter against a roughly $3.2 billion market cap that the sell-side is still modeling off FY26 print. But that’s not the point, the order book has already pulled into FY27, which is the period the street will be forced to re-cut models

around once the next two quarters of shipments and follow-on awards actually land.

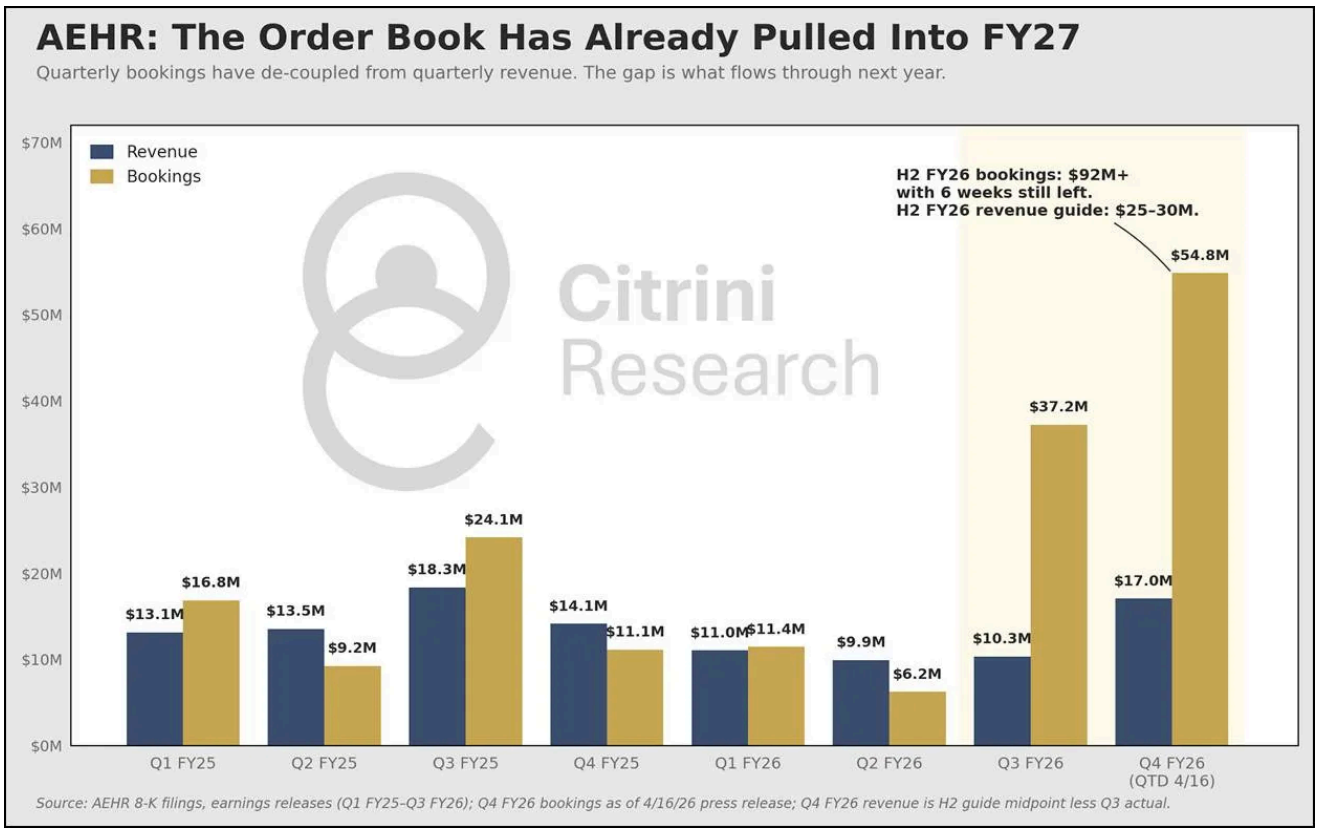
AEHR 报告称，其领先的超大规模 AI 客户（GOOG）已追加 4100 万美元的生产订单，用于定制 ASIC（TPU）的封装级老化测试，交付将于 2027 财年开始。这使得下半年的预订额超过 9200 万美元，且在本季度还剩六周的情况下，其市值约为 32 亿美元，而卖方分析师仍基于 2026 财年的数据进行建模。但这并不是重点，订单簿已经延伸到了 2027 财年，一旦未来两个季度的出货量和后续订单实际落地，华尔街将被迫围绕这一时期重新调整模型。

AEHR's FY27 revenue conversion turns them into a scarce piece of AI infrastructure with reorder cadence, installed-base consumables, second-customer optionality, and pricing power that scales with ASIC complexity.

AEHR 在 2027 财年的营收转型，使其成为稀缺的 AI 基础设施标的，具备订单复购节奏、装机量带来的耗材收入、第二客户的可选性，以及随 ASIC 复杂度提升而增长的定价权。

As custom ASICs get hotter and more mission-critical, early-life failures become an uptime liability sitting inside a training cluster that costs a LOT of money to run by the hour. Consistent with our entire memory testing thesis, burn-in turns into a cheap insurance policy on very expensive silicon. AEHR is being designed at both the package level (Sonoma) and the wafer level (FOX), and management has now pointed at silicon photonics. And the name still has SiC, GaN, flash, and HBM as the next demand vectors.

随着定制 ASIC 的发热量越来越高且任务关键性不断提升，早期失效（early-life failures）已成为训练集群正常运行时间的隐患，而这些集群每小时的运行成本极其高昂。正如我们整个存储测试论点所言，老化测试（burn-in）已成为昂贵芯片的一种廉价保险手段。AEHR 正在封装级（Sonoma）和晶圆级（FOX）进行设计布局，且管理层目前已指向硅光子领域。此外，该公司仍将 SiC、GaN、闪存和 HBM 视为接下来的需求增长点。



The customer-concentration argument that worked twelve months ago gets harder to defend every quarter. The recent $60 million ATM was dilutive in the most literal sense, but it cleared the balance-sheet question right as order flow is accelerating, which is unusual timing for one of these raises and a useful tell on management's read of demand. Put it all together and the asymmetry is clean. The discovery move has already happened. There is further re-rating yet to occur on their TPU socket and potential further custom ASIC wins from other hyperscalers.

十二个月前还奏效的客户集中度论点，现在每个季度都变得愈发难以立足。最近 6000 万美元的 ATM（市价增发）从最字面的意义上来说确实稀释了股权，但它在订单流加速之际解决了资产负债表问题。在此时点进行此类融资实属罕见，这也成为了解管理层对需求判断的一个有用信号。综合来看，这种不对称性非常清晰。探索性行情已经发生。其 TPU 插槽仍有进一步估值重修的空间，且未来有可能从其他超大规模云厂商处获得更多定制化 ASIC 订单。

We believe we can see AEHR double still despite the massive run up. We're also aware that it's likely to trade with significant volatility given crowding, so rather than

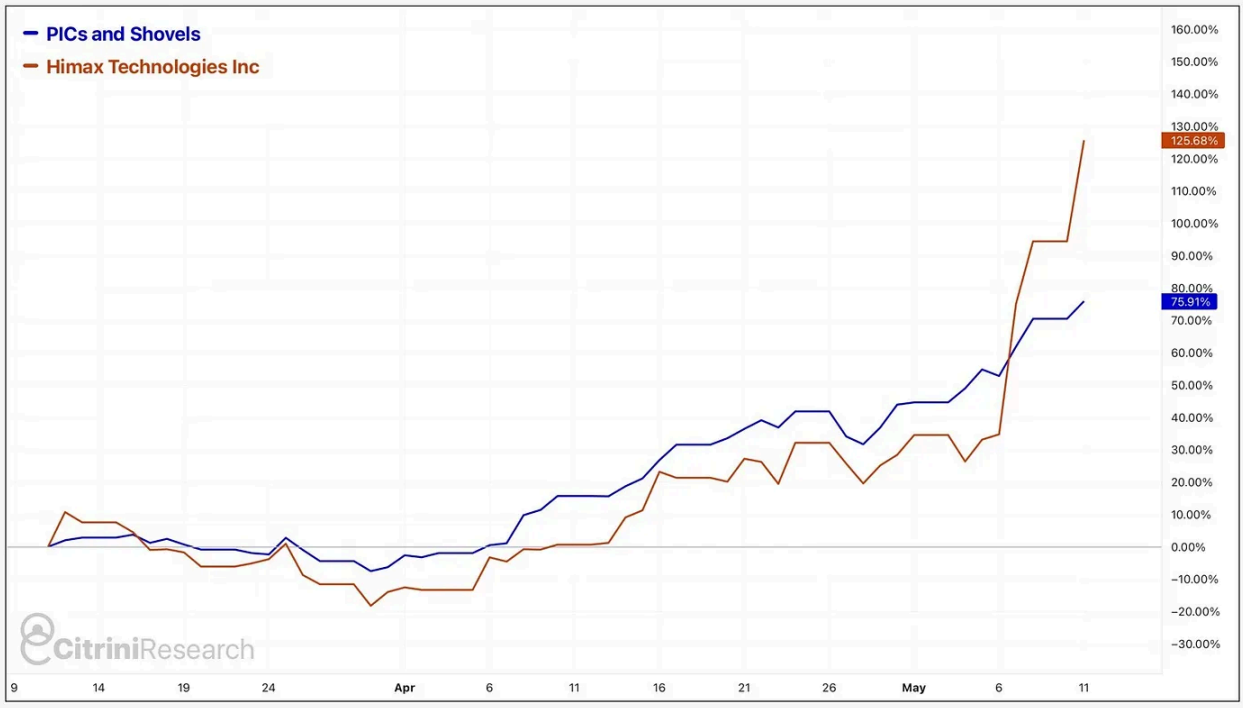
increasing here we'll be waiting to be a buyer on dips to increase our position.

我们认为，尽管 AEHR 已经经历了巨大的涨幅，但其股价仍有翻倍的潜力。我们也意识到，考虑到筹码过于集中，该股可能会出现剧烈波动，因此我们目前不会在此价位加仓，而是等待回调时再买入以扩大头寸。

Our second semis memo, "Let There Be Light", focused on the photonics supply chain. It has performed even more impressively, up more than 75% in just two months, while our single stock highlight **Himax** (**HIMX US**) soared +125% as earnings validated our thesis on their COUPE upside being significantly mispriced.

我们的第二份半导体备忘录《要有光》（Let There Be Light）重点关注了光子学供应链。其表现更为惊人，在短短两个月内上涨了 75% 以上，而我们重点推荐的单只股票 Himax (HIMX US) 则飙升了 +125%，因为财报证实了我们的论点，即市场对其 COUPE 业务的上涨空间存在严重误定价。

Our Photonics Basket and Single-Stock Pick HIMX Since Publication



Source: Citrini Research

We remain bullish on Himax as well. Just as we've done, we'll continue to update our active AI allocation opportunistically to some of our new picks and manage risk on our old ones. It should go without saying, but I'll still say it, that the move in semis over the past 6 weeks has been extremely violent and - even in a raging bull market - it's

likely for there to be pullbacks that provide better entry points on single stocks.

我们仍然看好 Himax。正如我们一直以来所做的那样，我们将继续根据时机灵活地将 AI 配置更新到一些新选股中，并对旧持仓进行风险管理。虽然这不言而喻，但我还是要强调，过去 6 周半导体板块的走势极其剧烈——即使是在狂热的牛市中，也可能会出现回调，从而为个股提供更好的入场点。

Some Thoughts on Where We're Headed

关于我们未来走向的一些思考

We are guilty of the very same ADHD focusing on the next thing as we have cautioned you against in the last section. While we were some of the earliest to look past the dismal environment for memory coming out of 2023 and emphasize the asymmetric HBM setup in memory OEMs back in January 2024, we also sold them too soon (in February 2026, right before they went parabolic).

我们同样犯了在上一节中告诫过你们的那种多动症式的错误，即过度关注下一件事。虽然我们是早在 2024 年 1 月就看透了 2023 年存储行业惨淡环境，并强调存储代工厂（OEM）中不对称的 HBM 布局的先行者之一，但我们也卖得太早了（在 2024 年 2 月，就在它们开始抛物线式暴涨之前）。

It has been somewhat difficult to balance, over the past 3 years (literally, May 29, 2026 will mark the 3 year anniversary of our first AI primer on the coming data center scale up that would 10x analyst estimates). Every time we think we're too bullish on AI, something surprises us more. Back in 2024, we wrote as part of our **January State of the Themes**:

在过去的 3 年里，保持平衡一直有些困难（确切地说，2026 年 5 月 29 日将是我们发布首篇关于即将到来的数据中心规模扩张 AI 入门报告的 3 周年纪念日，当时该报告预测的规模是分析师预期的 10 倍）。每当我们认为自己对 AI 过于乐观时，总会有一些事情带给我们更大的惊喜。早在 2024 年，我们在 1 月份的《主题现状报告》中写道：

“UBS is predicting AI industry revenue to grow to $420 billion by 2027, a 40% increase from its prior view of $300 billion.

“瑞银集团（UBS）预测，到2027 年，人工智能行业的收入将增长至4200 亿美元，较此前预测的3000 亿美元上调了40%。”

This would represent a 72% CAGR from $28 billion in 2022.

这意味着从2022 年的280 亿美元起，复合年增长率（CAGR）将达到72%。

And that’s conservative, by “bubble” standards.

按照“泡沫”标准来看，这还是保守的。

These are still easily achievable goals. If we are actually going into an AI bubble, it can get so much crazier. Looking at the telecom bubble of the early 2000s for a parallel, they were modeling all humans on earth with 2 cell phones. They modeled people in 2000 going from spending 15% of disposable income on internet access to 60%.

这些目标仍然很容易实现。如果我们真的进入了AI 泡沫，情况可能会变得疯狂得多。参考2000 年代初的电信泡沫，当时的模型假设地球上每个人都拥有2 部手机。他们预测2000 年的人们在互联网接入上的支出将从可支配收入的15% 增加到60%。

It is the base case that AI will be the size of the entire current global semiconductors market and 4x the annual capex of all the hyperscalers. This means there is still much room for estimates to get wildly out of hand in a bubble-like sequence.”

基准预测是，AI 的规模将达到目前全球半导体市场的总和，并且是所有超大规模云厂商年度资本支出的4 倍。这意味着，在类似泡沫的发展进程中，预测数据仍有巨大的空间变得极度失控。

Our view, back then, was that total AI spend would be five times higher than UBS’ estimate – the highest on the street at the time. And still, we were too conservative. It’s

now estimated that in 2027 AI spending will be in excess of $3 trillion.

当时我们的观点是，人工智能的总支出将比瑞银（UBS）的估计高出五倍——而瑞银在当时已是华尔街最乐观的预测。即便如此，我们还是过于保守了。据目前估计，2027 年人工智能领域的支出将超过 3 万亿美元。

What will someone laugh at today that will be true 24 months from now? That's the question that's been on the more bullish side of my thoughts.

今天人们会嘲笑什么，而在 24 个月后它却会成真？这是我思考中偏向看涨的一面一直在琢磨的问题。

Back then, it was primarily just me writing these newsletters and I had to try my best to temper my most bullish expectations because my job was to convince people – and we were still in a place where wild optimism got dismissed easily. I have always tried to balance a realistic outlook that takes into account the near certainty that estimates would get too insane, markets would overshoot to the upside and we'd blow and then burst a bubble with the deep conviction that AI is the most important investment theme of our lives. I think I've done a good job of that, but it has undoubtedly begun to be a hindrance.

那时候，主要是我一个人在写这些简报，我必须尽力克制自己最乐观的预期，因为我的工作是说服他人——而当时我们还处在一个盲目乐观很容易被嗤之以鼻的环境中。我一直试图在两种心态间寻求平衡：一方面是保持现实的展望，考虑到预测几乎肯定会变得过于疯狂、市场会超涨、我们会吹起并最终破灭一个泡沫；另一方面则是深信 AI 是我们一生中最重要的投资主题。我认为我在这一点上做得很好，但这无疑已经开始成为一种束缚。

Every cycle convinces the most rational of investors to become comfortable with expectations that are impossible. That's what the more paranoid side of my brain says.

每一个周期都会说服那些最理性的投资者，让他们安于那些根本不可能实现的预期。这是我大脑中多疑的一面所给出的想法。

So let's address what can happen from here…

那么，让我们来探讨一下接下来的走向……

On one side, there's the argument that we've entered a bubble. I think this is primarily driven by perceptions around the speed at which memory stocks have moved and the crowding in of capital that has made AI really the only thing that seems worth investing in. I am certain without a doubt that markets will get ahead of themselves and begin pricing in things that won't happen. Are we there yet, though?

一方面，有一种观点认为我们已经进入了泡沫期。我认为这种观点主要是由于人们感知到存储股的上涨速度过快，以及资本的过度拥挤，使得人工智能似乎成了唯一值得投资的领域。我深信市场无疑会过度透支预期，并开始对那些不会发生的事情进行定价。然而，我们现在已经到那个阶段了吗?

If one just looks at the rally in KOSPI - the forward PE multiple has gone DOWN since this rally began.

如果仅看韩国综合股价指数（KOSPI）的涨势——自本轮反弹开始以来，其远期市盈率（Forward PE）实际上是下降的。

We have gone from "memory is in a structural glut" to "memory isn't cyclical anymore" in the span of just more than two years. We still believe that memory, like all semiconductors, remains cyclical, despite AI. But the shape and duration of that cycle is highly dependent on the persisting bifurcation between HBM and commodity DRAM, and whether NAND follows a similar supply tightness from AI demand.

在短短两年多时间里，我们经历了从"存储芯片处于结构性过剩"到"存储芯片不再具有周期性"的转变。我们仍然认为，尽管有人工智能的推动，存储芯片与所有半导体产品一样，依然具有周期性。但这一周期的形态和持续时间，在很大程度上取决于 HBM 与商品化 DRAM 之间持续的分化，以及 NAND 是否会因 AI 需求而出现类似的供应紧张。

On the other side, there's the **leagues of serious experts talking about how insane the demand is.** They're all much smarter than I am.

另一方面，大批资深专家正在讨论这种需求是多么疯狂。他们都比我聪明得多。

The claim made here is that **by the early 2030s, global AI inference could be 10,000x larger, driven by 50x algorithmic progress, 50x hardware and systems progress, and**

4x data center capacity...

此处提出的观点是，到 2030 年代初期，全球 AI 推理规模可能会扩大 10,000 倍，这主要受算法进步 50 倍、硬件和系统进步 50 倍以及数据中心容量增长 4 倍的共同驱动……

CRA-I 2030 EXPECTATIONS

Agentic AI demand scenarios imply a multi-year supply bottleneck.

2030 P50 H100-EQUIVALENTS

Old chatbot-anchored assumptions vs new agentic-anchored assumptions.

SCENARIO	OLD (CHATBOT-ANCHORED)	NEW (AGENTIC-ANCHORED)
Acemoglu (bear bear)	0.2M	**8.9M**
Covello	0.2M	**12.8M**
Bear	0.4M	**22.8M**
Base	0.9M	**46.6M**
Bull	1.7M	**87M**
Jevons-max	2.0M	**106M**

SUPPLY CONSTRAINTS AT THE BASE CASE

At roughly 47M H100-equivalents of demand, several chokepoints bind.

CONSTRAINT	2030 CEILING	DEMAND AT 47M H100-EQ BASE
TSMC CoWoS (incl. OSAT)	~5M wafer-equivs/yr	15M Rubin × ~2× wafer area = **~30M wafer-equivs of demand**
HBM4 capacity	300 PB/yr	15M Rubin × 288 GB = **~4.3 EB just for chips**
Transformer (datacenter power infra)	3-yr lead time	135 GW @ 2030 needs orders placed **now** (2027 lead, 2030 online)

TAKEAWAY

In the base case, demand reaches roughly 47M H100-equivalents by 2030, while key supply chokepoints sit far below that level. If this base case is right, demand likely outruns supply for

more than 5 years.

Source: CRA-I expectations summary. Design: Citrini Research.

Citrini Research

If expectations like this become the consensus, then yes, it's not difficult to envision a scenario where analysts can easily justify SNDK or MU or SK Hynix tripling from here. And maybe Sandisk does continue to see datacenter mix shift, we normalize peak earnings, value it at a 5% FCF yield, and it trades multiples higher from here. Or, HBM demand continues to voraciously consume any and all incremental capacity out to 2029, putting SK Hynix at 2x 2029E EPS.

如果这类预期成为共识，那么确实不难想象分析师们可以轻而易举地为 SNDK、MU 或 SK Hynix 在现有基础上翻三倍找到合理依据。也许 Sandisk 的数据中心业务占比会持续提升，我们将其峰值收益标准化，并按 5% 的自由现金流收益率（FCF yield）进行估值，其股价可能会在目前基础上再涨数倍。又或者，HBM 的需求将持续贪婪地消耗直到 2029 年的所有新增产能，从而使 SK Hynix 的股价达到 2029 年预期每股收益（EPS）的 2 倍。

These outcomes are entirely possible because they're much more a function of imagination about the future than discounting an extrapolation of the present. That imagination has repeatedly proven to be too conservative in its optimism. But as significant as this all is, I think it's worthwhile to reserve a small part of your brain to remember to be on the lookout for the signs.

这些结果完全是有可能实现的，因为它们更多地取决于对未来的想象，而非对现状外推的折现。事实一再证明，这种想象在乐观程度上往往过于保守。尽管这一切意义重大，但我认为仍值得在脑海中保留一小部分空间，提醒自己时刻留意那些迹象。

I'm not too concerned about whether or not we are entering a bubble, because every bubble that's burst in the past 50 years has happened following the Fed tightening monetary conditions. Right now, the economy is great and the Fed is on hold. It's not going to be the technology that causes this to stop because we can keep scaling it and keep spending and keep coming up with new Jevons-y reasons why we will never have enough compute.

我不怎么担心我们是否正在进入泡沫，因为在过去 50 年里，每一次泡沫的破裂都发生在美联储收紧货币条件之后。眼下，经济状况良好，美联储正处于观望状态。技术本身不会导致这一进程停止，因为我们可以不断扩大规模，持续投入资金，并不断提出符合“杰文斯悖论”的新理由，来说明为什么我们永远无法拥有足够的算力。

Like always, it will be macro. So, as uninspiring as this is, my strategy remains the same as always – play the game, remain on the side of the technology, don't anesthetize myself with unrealized gains to the point where I ignore the indications that have consistently signaled the top dozens of times before or buy in to the idea that cycles don't exist anymore. Keep updating the potential list of scenarios we can encounter that pose risks to ever increasing estimates and maintain a robust set of signals that indicate we're going down one path or another.

一如既往，这将取决于宏观环境。因此，尽管这听起来平淡无奇，我的策略依然如故——参与其中，站在技术进步的一方，但不要被未实现收益冲昏头脑，以至于忽视那些曾数十次准确预示顶部的信号，也不要盲目相信周期已不复存在的说法。持续更新我们可能面临的潜在风险情景列表（这些风险可能威胁到不断攀升的预测值），并维持一套稳健的信号体系，以指示我们正走向何种路径。

This article is for informational purposes only and does not constitute investment advice. By accessing this material, you agree to our Terms of Service.

本文仅供参考，不构成投资建议。访问此材料即表示您同意我们的服务条款。

Recommend Citrini Research to your readers

向您的读者推荐 Citrini Research

Citrini Research provides insights on thematic equity investing and global macro trading—with cross-asset, lateral thinking. Our promise: you'll never have to ask "what's the trade?"

Citrini Research 提供关于主题型股票投资和全球宏观交易的见解——具备跨资产的横向思维。我们的承诺是：您永远不必再问"交易机会在哪里？"

Recommend 推荐



157 Likes 157 次赞 · 4 Restacks 4 次加仓

Discussion about this post

关于此帖的讨论

Comments 评论 Restacks 转贴



Write a comment...



seanpm 7m

“I’m thinking we might write about a theme that isn’t a derivative of AI this week if that’s ok.”

LIKE REPLY SHARE

© 2026 Citrinitas Capital Management Inc. · Publisher Privacy · Publisher Terms
Substack · Privacy · Terms · Collection notice
Substack is the home for great culture