

权重刻进硅之后，卡点搬去了掩模房

AI 推理硬件 · 固化推理硅（weights-in-silicon）路线的卡点归属

把权重刻进硅解决的是物理问题，却制造了一个时间问题——真正的卡点不在芯片，在掩模周转

涌现资本产业研究部 · 2026-08-15

命题：固化推理硅把「内存墙」换成了「时间墙」。谁能把一版权重在两个月内变成可交付的硅，谁就握住这条路线的咽喉——而这个能力不在芯片公司手里。

一、30秒结论卡

一句话判断：AMD 收购 Taalas 与 NVIDIA 收编 Groq，在八个月内把「固化/驻留权重」这条路线从独立赛道变成了两大 GPU 厂的机架内部件；**这条路线的物理逻辑成立、独立投资逻辑不成立**，可捕获价值沉到成熟节点代工与掩模周转服务这一层。

机会状态：● 观察为主。固化推理硅本身**不构成新的可投卡点**；它改变的是既有卡点的相对硬度排序。

市场错在哪：市场把这笔收购读成「AMD 拿到了挑战 NVIDIA 推理的武器」。真正的读数是另外两条——第一，NVIDIA 在 2025 年底花 200 亿美元收编 Groq、并在 2026 年 3 月把自研的 prefill 专用 GPU (Rubin CPX) 移出路线图，等于用行动承认 **decode 阶段需要非 GPU 架构**；AMD 这笔收购是同一判断的第二次确认，不是差异化。第二，固化路线的经济性完全押在一个工艺技巧上——**只改顶层两层金属掩模、两个月完成权重到硅**。这条流水线的速度决定生意成不成立，而它的物理执行方是代工厂与掩模房，不是设计公司。

重点标的：TSM ● · AVGO ● · NVDA ● · AMD ● · MU/SNDK ● (HBM 侧未被实质削弱)

最大反证：若 AMD 交割后首个客户案例给出可核实的 \$/百万 token 优势，并证明两个月周转已在量产中跑通，则固化硅从「niche 加速器」升格为通用推理的成本基准，届时受损方从「无」变为「HBM 与可编程 ASIC 双线」，本篇分层需整体重写。

二、投资要点

框架：decode 是内存受限的，所以所有路线都在回答同一个问题——怎么让「权重到计算单元」这段距离消失。固化派让分母归零，代价是把芯片寿命绑死在模型寿命上。

观点	方向	依据
权重固化的物理收益真实，商业窗口极窄	●	主力模型换代中位 4.1 个月，全套掩模流片 6–12 个月，窗口为负
这条路线已无独立上市纯 play	●	Taalas 并入 AMD、Groq 核心 IP 与团队并入 NVIDIA
真正被加强的是成熟节点代工与掩模周转	●	HC1 用 N6、LPX 用三星 4nm、Sohu 用 N4P——全部不在最先进节点
HBM 需求未被实质替代	●	prefill 与 KV cache 仍在 HBM 侧；SRAM 密度已停止随节点缩小
可编程 ASIC 仍是吃掉推理量的主力	●	超大厂自研既是设计方又是唯一客户，无需说服市场

三、钱流：谁在为 decode 付钱

这条路线上迄今发生的现金与股权流动，比任何性能宣称都更能说明产业判断。

流向	金额	时点	读数
NVIDIA → Groq（授权加团队并入）	约 200 亿美元	2025–12–24	为 decode 专用架构付出的最高价
AMD → Taalas	未披露	2026–08–06 宣布	同一判断的第二次确认
一级市场 → Etched	累计逾 11 亿美元	至 2026–07	固化架构路线的独立融资上限
一级市场 → Taalas	累计 2.19 亿美元	2024 / 2026–02	权重固化路线并入前的全部投入
一级市场 → d-Matrix	C 轮 2.75 亿美元	2025–11	存内计算路线

价格侧唯一的公开锚：NVIDIA 为 Groq 3 LPX 机架给出的目标价格点为每百万 token 45 美元。这是全行业目前唯一可核实的固化/驻留路线成本读数，其余各家均无公开的每百万 token 或每瓦 token 数据。

这张表的读法：两家 GPU 巨头在八个月内先后为同一条技术路线付钱，而独立玩家的融资量级（十亿美元级）与部署这条路线所需的资本量级（数百亿美元级）之间相差两个数量级。**资本流向本身就是「这条路线不会有独立赢家」的最强证据。**

四、事件与技术：被刻进硅的到底是什么

AMD 于 2026–08–06 宣布与 Taalas 达成最终协议，**对价未披露**，需通过惯例交割条件与监管批准，市场预期 2026 年第四季度交割。Taalas 2023 年成立于多伦多，累计融资约 2.19 亿美元（2024 年 5,000 万

美元 + 2026 年 2 月 1.69 亿美元)，创始人 Ljubisa Bajic 为 Tenstorrent 联合创始人，更早在 AMD 与 NVIDIA 做架构。

技术实现的关键不是「刻进硅」这个说法，而是三个具体机制：

1. **权重存在掩模 ROM 里**。芯片上有两套 recall fabric：mask ROM 承载不变的权重，SRAM 承载 KV cache 与微调 adapter。前向传播过程中**不存在从外部存储读权重这个动作**，HBM 因此被整体拿掉。
2. **单个晶体管存 4 bit 权重并完成该次乘法**。这是密度的来源，也是精度的代价——Taalas 自认相对 GPU 基准存在质量退化。
3. **换模型只改约 2 层金属掩模**（整套约 100 层掩模中的顶层金属），厂商声称权重到可交付 PCIe 卡约 2 个月，设计生成约 1 周。

首颗验证芯片 HC1：TSMC N6、815 mm²（逼近光罩上限）、530 亿晶体管、单片承载 80 亿参数，单卡 250W，十卡装进 2.5kW 风冷双路服务器。下一代 HC2 目标单片最高 200 亿参数。承载万亿参数模型的路径是 50 颗芯片流水线并行——**这一步把内存墙换成了互连墙**。

性能口径必须钉死：**17,000 tokens/s 是 Llama 3.1 8B、1K 输入/1K 输出、单用户单流读数**，其对照的 H200 约 230 tokens/s 同为单流。GPU 的商业部署跑高批量，聚合吞吐口径下的差距远小于此。厂商另有「1000 倍性能功耗比」的表述，属总括宣称，无第三方复现，**不可作为任何估值输入**。

五、第一过滤器：它卡谁的脖子

把「模型主力寿命」与「权重到硅的交付周期」放在同一根时间轴上，这条路线的成败是一道减法。

主力模型寿命（自建可复现口径 = 前一版发布日 → 同族主力后继版发布日）：取 2024-05 至 2026-04 的 11 个真实换代样本（Llama 3.1→3.3→4、Qwen2.5→3、DeepSeek V3→R1→V3.1→V3.2、GPT-4o→4.5→4.1→5→5.2→5.5），**中位数 4.1 个月**；只留大版本代际口径为 4.4 个月，两套口径同向。

可服役窗口 = 模型寿命 - 交付周期：

路径	交付周期	窗口	判定
全新流片（全套掩模·先进节点）	6-12 个月	-1.9 至 -7.9 个月	● 芯片到货即过期
顶层金属改版（厂商声称）	2 个月	+2.1 个月	● 勉强为正，无缓冲
权重冻结的窄任务（语音/翻译/嵌入/边缘）	6-12 个月	+12 至 +30 个月	● 唯一舒适区

这张表是本篇的地基，它给出三个结论：

其一，任何需要重做全套掩模的固化路线，在前沿模型上数学上不成立——芯片下线时，被固化的那一版权重已被后继版替换过至少一轮。

其二，整条路线的商业模型押在一个工艺技巧上：半定制基底晶圆 + 只换顶层金属。这不是「芯片好不好」的问题，是「流水线快不快」的问题。**这类公司在投研上应当按掩模周转服务商来估，而不是按芯片设计公司来估**——而周转速度的物理执行方是代工厂与掩模房。

其三，真正的目标客户不是追前沿的人。企业固化自己的微调模型、愿意把权重冻结 12—24 个月的负载，才是窗口舒适的那一段；权重多年不动的窄任务（语音识别与声纹、机器翻译、嵌入与向量检索、边缘小模型）合计约 150—300 亿美元的软件与服务口径规模，硅片可捕获份额远小于此。固化推理即使全胜，也是一个 niche，除非它成功迁移到「企业固化自有微调模型」这个尚未被数据证实的大市场。

六、格局：两条固化路线，八个月内各归其主

2025 年底至今，这个赛道发生了一次彻底的收编。

NVIDIA 侧：2025-12-24 以约 200 亿美元完成对 Groq 的授权加团队并入（法律上非整体收购，以规避审查周期与反垄断风险，创始人与核心工程团队并入）。产物是 2026 年 GTC 发布的 **Groq 3 LPX**——三星 4nm 制造，单 die 512 MB 片上 SRAM、150 TB/s 带宽，明确定位为 **decode 阶段专用协处理器**，Rubin GPU 管 prefill，LPU 管 token 生成，计划 2026 年第三季度出货。NVIDIA 给出的口径是：配 Vera Rubin NVL72，在万亿参数模型上每兆瓦吞吐较 Blackwell NVL72 提升至多 35 倍，**目标价格点每百万 token 45 美元**。同期，2025 年 9 月发布的 prefill 专用、GDDR7 路线的 Rubin CPX 于 2026 年 3 月被移出路线图。

半年之内，NVIDIA 把 prefill/decode 解耦的介质选择从「GDDR7 做 prefill」翻转成「SRAM 做 decode」。翻转本身就是判断：**真正的钱在 decode 侧**。

AMD 侧：收购 Taalas，官方口径是并入加速器路线图、与 Instinct GPU 做系统级方案，与 Helios 机架、EPYC、ROCm 并列全栈。**AMD 通稿并未表述「GPU 做 prefill、Taalas 做 decode」的分工**，这是产业侧的合理推演，本篇按推演处理，不作事实陈述。

剩下的人：Etched 走的是固化架构而非固化权重（TSMC N4P，transformer 专用电路），累计融资逾 11 亿美元、签约合同逾 10 亿美元，但合同对方未具名，且截至 2026 年中未量产、无第三方验证。d-Matrix 走数字存内计算，Corsair 于 TSMC 6nm 量产、2026 年 6 月起发货，具名客户未披露。Cerebras 走晶圆级 SRAM 并已在纳斯达克上市，技术侧已被验证，其核心风险不在架构而在客户集中度。

这张地图的读法：SRAM 与固化路线赢了架构辩论，输了独立性。对投资人而言，**这条路线上已经不存在「买到纯正固化推理」这个选项**——它变成了 NVDA 与 AMD 机架里的一个部件，其价值被稀释进两家万亿与千亿级市值的整体叙事中。

七、价值链五层与受益链：哪一层被加强，哪一层被削弱

需求源：超大厂 + 前沿实验室的推理负载 (decode 段)
↓
一级供应商：NVDA (GPU+LPX 全机架) · AMD (Instinct+Taalas) · 超大厂自研 ASIC
↓
真卡口：成熟节点先进产能 + 掩模制作与周转 + 先进封装
↓
可投代理：TSM · AVGO · 三星 (非美股直投)
↓
已拥挤 / 未定价：AMD 前瞻估值已计入 2027 兑现 / 掩模周转能力尚未被单独定价

层	变化	方向	依据
成熟节点代工 (N6/N4P/4nm)	加强	●	三条固化路线全部避开最先进节点，落在成熟高良率节点
掩模制作与周转	加强 (本篇新增卡点)	●	2 个月周转是整条路线的生死线，能力集中在极少数产线
可编程 ASIC 集成	维持	●	超大厂自研既是设计方又是唯一客户，无需说服市场
HBM	未被实质削弱	●	prefill 与 KV cache 仍在 HBM 侧；SRAM 密度停滞使「装下大模型」有物理上限
内存接口与网络	中性偏正	●	多芯片流水线并行把内存墙转成互连墙，互连需求上升
纯固化硅概念	消失	●	两家最激进玩家分别被 AMD、NVIDIA 收编

关于 HBM 的判断需要单独说清，因为这是本篇最容易被误读的一层。decode 卸载是本轮唯一在结构上指向「少用 HBM」的方向，但三条物理约束限制了它的上限：其一，prefill 阶段仍是计算受限、仍跑在配 HBM 的 GPU 上；其二，KV cache 随上下文长度线性增长，固化芯片的 SRAM 侧正是用来装它的，容量约束反而更紧；其三，SRAM 位密度已基本停止随节点缩小，「把更大的模型放进片上」不再是一个可以靠制程解决的问题。因此本篇不给 HBM 层做任何减分——它面对的是量的边际分流，不是定价权的转移。

受益链拆解 (需求源 = NVIDIA / AMD 的 decode 侧机架)：

- 直接受益：TSM (三条固化路线与主要 ASIC 的共同代工方) · AVGO (可编程 ASIC 集成方)
- 二阶受益：先进封装与测试环节 (多芯片流水线并行使单系统芯片数量上升) · 互连与网络 (内存墙转互连墙)
- 隐形受益：掩模制作与快速改版服务 (尚未被单独定价的一层，无纯正上市代理)
- 伪受益：一切以「AI 推理概念」自我标签、但与上述两家机架无实际供货关系的标的
- 可投代理：TSM · AVGO · NVDA · AMD
- 不可投但需监控：Taalas (待并入 AMD) · Etched (私有) · d-Matrix (私有) · 三星 (LPX 代工方，非美股直投)

八、公司卡

TSM · \$426.35 · PE 32.2 · 市值约 2.21 万亿美元 ●

位置：三条固化路线（HC1 的 N6、Sohu 的 N4P、Corsair 的 6nm）与几乎全部可编程 ASIC 的共同物理执行方。为什么重要：固化路线的胜负手是周转速度，而周转速度是代工厂的工程能力，不是设计公司的。证据：HC1 明确采用 TSMC N6 并使用其快速流片工作流。估值状态：32 倍 TTM，是本篇标的中唯一估值与增速不构成张力的。催化：AMD 交割后首个 weights-to-silicon 客户案例的节点选择。反证：若固化路线始终停留在样片阶段，这条增量对 TSM 体量而言不具材料性。

AVGO · \$392.99 · PE 65.3 · 市值约 1.87 万亿美元 ●

位置：可编程 ASIC 路线的共同集成方。为什么重要：无论固化路线成败，2026 年推理量的主力仍由超大厂自研 ASIC 承担，而这条路线的收费站是物理设计与集成。反证：超大厂自研深化到自建设计团队，收费站被绕过。

AMD · \$514.39 · TTM PE 约 132 倍 (GAAP) · 前瞻约 63—67 倍 · 市值约 8,390 亿美元 ●

位置：拿到权重固化路线，但对价未披露、交割未完成、无公开时间表。基本面：FY26 Q2（截止 2026-06-27）营收 115.36 亿美元同比增 50%，数据中心 67.18 亿美元同比增 107%，GAAP/non-GAAP 毛利率 54%/56%，摊薄 EPS 1.38/1.66 美元，Q3 指引 130 亿美元上下 3 亿。为什么是 ● 而非 ●：OpenAI 与 Meta 各 6 GW 的协议分别附带至多 1.60 亿股、行权价 0.01 美元的绩效认股权证，合计 3.20 亿股，相当于现有股本约 19.6%（按行权后股本口径为 16.4%）——这是用股权补贴 GPU 售价，且 1 GW 首批交付要到 2026 年下半年才开始。按「已发生的小收入 > 承诺的大加速」，本次收购与两份 GW 级协议均不构成潜力上调依据。反证：交割后首个客户案例给出可核实的成本优势。

NVDA · \$225.16 · PE 34.5 · 市值约 5.45 万亿美元 ●

位置：已在同一机架内同时握住 prefill 与 decode 两端。为什么重要：LPX 是目前唯一给出公开价格点（每百万 token 45 美元目标）的 decode 专用硅，这个数字将成为整条路线的成本基准线。反证：Q3 出货时点滑期，或 35 倍每兆瓦吞吐在第三方复现中大幅缩水。

MU / SNDK 等 HBM 与存储层 ●

维持中性。decode 卸载是唯一结构性利空方向，但受 prefill、KV cache 与 SRAM 密度停滞三重约束，2026—2028 年内不构成需求侧的一阶变量。

九、数据分歧与口径

1. 「固化硅便宜多少」目前没有一手依据。全行业公开的 \$/百万 token 数据仅有 NVIDIA 为 LPX 给出的 45 美元目标价格点一个。Taalas、Etched、d-Matrix 三家均无可核实的 tokens/s/W 或 \$/百万 token。所有「便宜 N 倍」的说法在这个数字被填上之前都不成立。
2. 顶层金属改版的单次成本未知，这是整个命题的核心空洞。行业公开的掩模成本口径本身也分歧显著：16/14nm 一套约 500 万美元、7nm 约 1,500 万美元是一条线；先进节点掩模组推向 4,000 万美元量

级是另一条线。两条线并存，本篇取量级参考而不入定量结论。

- 3. 跨厂商性能宣称不可排序。17,000 tokens/s（单用户单流）、Sohu 相对 H100 约 20 倍、Corsair 30,000 tokens/s（聚合口径）、LPX 每兆瓦 35 倍——模型规模、批量、精度、分母单位全不相同，任何横向排名都是口径事故。
- 4. 2026 年超大厂自研 ASIC 出货约 190 万颗为第三方估算，无厂商披露背书，仅作量级参考，不作定性结论的论据。
- 5. NVIDIA—Groq 交易定性：CNBC 口径为收购资产，Groq 口径为非独家授权，200 亿美元金额有多家一手媒体（Tom's Hardware、The Next Platform、Forbes）与创始人本人受访确认，但无监管文件。本篇按「授权加团队并入」表述。
- 6. Rubin CPX 表述为「已移出路线图」而非「已取消」——NVIDIA 未发官方声明。
- 7. AMD 两个股数基数并存：市值口径约 16.32 亿股、EPS 反推加权摊薄约 16.64 亿股，稀释比例的分子分母已在第七节标明各自口径。TTM PE（GAAP）与前瞻 PE（卖方一致预期，多为 non-GAAP）不可直接比较。

十、情景

情景	核心假设	结果	标的动作
🔴 熊（概率偏高）	两个月周转仅停留在设计流程理论值，首个客户案例延后或无具名客户	固化路线退回窄任务 niche，AMD 该笔收购不产生可辨识收入	AMD 维持 🟡 不加分；TSM 该增量不具材料性
🟡 基准	周转跑通但客户集中在企业自有微调模型，规模停在数十亿美元级	成为 GPU 机架里的一个可选部件，价值捕获归代工与整机	TSM 🟢 维持；AVGO 🟢 维持；AMD/NVDA 中性
🟢 牛	有可核实的每百万 token 成本优于 45 美元，且周转在两个月内量产验证	固化硅成为通用推理成本基准，HBM 与可编程 ASIC 双线承压	第七节分层整体重写，HBM 层转 🔴

十一、风险与反证

风险	会推翻什么	观察点
两个月周转从未在量产中跑通	推翻整条固化路线的商业模型（窗口回到负值）	AMD 交割后首个客户案例的实际交付周期
企业固化自有微调模型的需求不存在	推翻 niche 之外的所有增长空间	首批具名企业客户与其权重冻结时长
\$/百万 token 优势不成立	推翻「固化更便宜」这一全部前提	对照 LPX 的 45 美元目标价格点
SRAM 容量天花板被新工艺突破	反向推翻本篇对 HBM 的中性判断，HBM 需转 ●	3D SRAM 堆叠或新型片上存储的量产公告
AMD 两份 GW 级协议的权证条款超预期稀释	推翻前瞻 63–67 倍这一读数的可用性	下一份 10–Q 中权证的会计处理与摊薄计入
固化路线被超大厂直接内化	推翻 AMD/NVDA 在这一层的价值捕获	超大厂自研 ASIC 中出现权重固化设计

十二、行动清单

重点观察：AMD 交割后首个 weights-to-silicon 客户案例——实测 tokens/s/W、定价、以及所选代工节点。这一个案例同时验证本篇的三个空洞（周转、成本、客户）。

只监控不买：一切以「固化推理」「刻进硅」为标签的概念标的。这条路线已无上市纯 play，任何声称纯正敞口的都需先核实其与 AMD/NVIDIA 机架的实际关系。

绝不追：把这笔收购当作 AMD 追赶 NVIDIA 的证据而追高。NVIDIA 早八个月做了同一判断，且已有出货时点与价格点。

下次看这五个数据：① AMD 交割公告与对价披露；② Groq 3 LPX 第三季度实际出货与第三方 benchmark；③ HC2 是否兑现单片 200 亿参数；④ 任一家给出可核实的 \$/百万 token；⑤ 顶层金属改版的单次掩模成本。

信号出现要重写结论：任一固化芯片在公开可复现的基准上给出优于 45 美元每百万 token 的成本，同时证明周转在两个月以内——届时受损方从「无」变为「HBM 与可编程 ASIC 双线」，第六节分层整体重写。

十三、来源置信度与数字三道关自评

结论 / 数字	来源	日期	口径	置信度
收购成立、对价未披露、需监管批准	AMD IR 通稿	2026-08-06	一手	A
AMD FY26 Q2 财务与 Q3 指引	AMD IR 稿 + 10-Q	2026-08-04	一手	A
OpenAI / Meta 各 6 GW 与认股权证	AMD 8-K、AMD IR 稿	2025-10 / 2026-02	一手	A
HC1 规格（N6/815mm ² /530 亿晶体管/80 亿参数）	Taalas 产品页 + The Next Platform	2026-02	厂商 + 产业媒体	B
权重载体为掩模 ROM、改 2 层金属、2 个月周转	The Next Platform、The Register	2026-02 / 08	厂商声明	C
17,000 tokens/s（单用户单流）	Taalas 产品页	2026	厂商基准，无第三方复现	C
Groq 3 LPX 规格、Q3 出货、45 美元/百万 token	NVIDIA 产品页 + Tom's Hardware + The Next Platform	2026-03	一手 + 多源 二手	B
NVIDIA-Groq 约 200 亿美元	Tom's Hardware、The Next Platform、Forbes（含创始人受访）	2025-12 / 2026-03	多源，无监管文件	B
模型换代中位 4.1 个月	涌现资本自建口径，11 个换代样本	2024-05 至 2026-04	自建，公式见第四节	B
掩模成本量级	IBS、SemiAnalysis 等	—	口径分歧，量级参考	C
2026 自研 ASIC 出货约 190 万颗	第三方估算	2026	无厂商背书	D
现价 / 市值 / PE	实时行情接口	2026-08-15 02:06 UTC	收盘	A

数字三道关自评

- **量级一致性：**AMD FY26 Q2 分部加总 67.18 + 38.41 + 9.77 = 115.36 亿美元，等于总营收；EPS 与净利反推的摊薄股数（GAAP 16.64 亿股 / non-GAAP 16.63 亿股）一致；Q3 指引 130 亿美元对应环比约 13%，与同比约 41% 自洽。Groq 3 LPX 机架：256 颗 × 512 MB ≈ 128 GB、256 × 150 TB/s ≈ 38.4 PB/s，与官方机架级读数自洽。
- **单位显式化：**全文美元与「亿/万亿美元」逐项标注；tokens/s 一律标明单流或聚合分母；掩模成本区分「单套」与「掩模组」。

- **两源以上**：AMD 财务与协议条款走公司公告与 10-Q；Taalas 技术规格为厂商口径加两家产业媒体；NVIDIA—Groq 金额为多家一手媒体加创始人受访，无监管文件，已在第九节单列。

涌现资本产业研究部 · 2026-08-15

免责声明：本报告为产业研究与信息聚合，不构成任何证券的买卖建议。所标注的方向评级为产业逻辑判断，非投资评级。文中厂商性能声明均未经第三方独立验证，不可作为估值输入。投资者应独立判断并自担风险。*